

趙民德教授 詩文選集



趙民德教授於 2026 年 8 月 15 日辭世，
謹以此一詩文選集紀念先生

目錄

生平簡介.....	6
編者說明.....	6
卷一 臉書隨筆.....	7
一、關於羅小華教授和我(2025.11.5).....	8
二、危機與心態（短語）(2025.10.10).....	8
三、生日只能有蛋糕嗎——水調歌頭(2025.10.6).....	9
四、虞美人——賀南區 34 屆統計研討會(2025.6.18).....	10
五、根號負一(2025.6.18).....	11
六、青年守則第 13 條(2025.6.9).....	13
七、苦難的歲月(2025.5.18).....	13
八、最長詞牌很難填嗎——鶯啼序(2025.5.16).....	14
九、七絕二首（忠孝醫院即興）(2025.5.8).....	15
十、拜年(2025.9.4).....	16
十一、江河旋律(2025.7.5).....	16
十二、寒門崛起(2025.5.26).....	17
十三、夢微之(2025.4.27).....	17
十四、中藥也在找搭檔(2025.4.23).....	18
十五、我領堅振(2025.4.19).....	18
十六、集句(2024.4.27).....	19
十七、年會(2023).....	20
十八、敬悼黃文璋教授（2023.11.21）.....	20
十九、對聯一副(2022.2.1).....	20
二十、賀友鐘兄新著(1995.6.1).....	21
二十一、轉貼兩則.....	22
卷二 詩詞卷.....	23
一、過沈從文宅(2005.8.8).....	24
二、某一個晚上的小詩(2006.2.1).....	24
三、江湖小隱(2006.6.6).....	24
四、武漢三鎮(2006.8.8).....	25
五、蒼狗(2006.9.9).....	25
六、2006 歲暮雜感(2006.12.12).....	25
七、給〈北風〉的歌者(2007.2.13).....	26
八、念奴嬌(2007.9.24).....	26
九、過柳永祠(2007.9.24).....	27
十、八聲甘州(2007.9.24).....	28
十一、子陵釣台(2007.9.24).....	28

十二、今古四喜詩(2010.3.3).....	28
十三、生查子(1964.2.1).....	29
十四、秋興(1964.2.1).....	29
十五、舊作(1964.2.1).....	29
十六、地震之後的迴想(1999.2.9).....	30
十七、北宜公路十六點二公里(2002.2.7).....	31
十八、某秋日登山作(2003.2.1).....	31
十九、四不詩(1981).....	32
卷三 散文卷.....	33
一、舊函(2002.2.6).....	34
二、委托書的世界(2003.2.1).....	35
三、關於蔡文甫的《天生的凡夫俗子》(2004.2.1).....	36
四、小言(2004.2.2).....	37
五、仲秋時節好讀詩(2004.2.7).....	38
六、年輪(2006.2.7).....	41
七、一壺濁酒喜相逢(2007.2.3 轉貼).....	42
八、寫書是一種情結(2007.2.3).....	44
九、羞愧與感恩(2007.2.9).....	45
十、Old Letter(2001.12.12).....	50
十一、義大利心影(1995.10.10).....	51
十二、山中日記(1968.2.2).....	60
十三、同學重聚記事(2010.2.1).....	65
十四、破而後立(2010.3.4).....	65
十五、不夠智慧時不如不要智慧(2010.3.3).....	66
十六、銀髮再學中文輸入(2010.3.3).....	67
十七、如夢的夢德里(1964.2.3).....	68
十八、靜靜的明湖(1967.7.7).....	69
十九、南港微雨(1993.6.2).....	70
二十、回首來時路(2000.10.2).....	71
二十一、今夜星辰依然閃爍(2001.11.27).....	77
二十二、Goddamn pandas(2007.2.4 轉貼).....	79
卷四 學術隨筆.....	80
一、我所知道的一點 Data Mining (2001.2.2).....	81
二、關於東華本《統計學》(2007.2.9).....	86
三、把春波都釀作一江春酎(1999.2.2).....	87
四、昔年種柳(1982).....	91
五、中文學術論文的規格芻議(數學傳播,十六卷二期).....	95
六、統計已死，統計萬歲！(1999.7.15).....	101

七、品味、執著和基本能力(數學傳播，二十七卷四期).....	113
八、萬物有常·世事多變(2001).....	124
九、統計是數據科學.....	130
十、關於統計論文的撰寫.....	136
附錄 作者自述.....	260
個人簡介.....	261
出生日期.....	261
關於酒店主人.....	261
關於部落格「趙家酒店」.....	261

生平簡介

趙民德教授以第一志願進入台大數學系就讀（1957），後來在美國加州大學柏克萊分校的統計系取得博士學位（1967）。畢業後在貝爾實驗室及 AT&T 公司服務，開啟他分析數據資料進行統計研究的工作。

在美國的工作期間內，曾回台於中研院數學所客座，實質參與推展國內統計的發展。後來更辭去美國的工作，舉家回台，擔任中央研究院統計科學研究所籌備處主任及成立後第一任及第二任所長（1982-1993）。趙先生為初創的中研院統計所廣徵專家學者及培育年輕人才，並且以身作則，創造一種自由且熱忱研究的氣氛，為國內的統計發展開啟一個新的局面，並將台灣統計帶向世界級的領域。

趙教授退休後仍筆耕不輟，於部落格「趙家酒店」¹及臉書上留下大量詩詞、散文與學術隨筆，記錄其學思歷程、生活感懷與對故友師長的追憶。

編者說明

本集彙整趙民德教授歷年發表於臉書及部落格「趙家酒店」¹之詩詞、散文與隨筆，以及趙民德教授臉書之隨筆，依性質分為「臉書隨筆」、「詩詞卷」、「散文卷」及「學術隨筆」四部，並附生平簡介於卷首、作者自述於卷末。原文散見各處、多有重複刊載，今去其重複、依寫作或發表年代大致排序，以便閱覽，並以此紀念先生。

¹ 原存網站已遺失

卷一 臉書隨筆

一、關於羅小華教授和我(2025.11.5)

這一陣台灣統計界的大事是羅小華教授的去世。他是和我同一位博導的師弟，中研院的統計所從籌備處的時代起，他就幾乎年年都來客座。記得 1980 時我就和他來往：他在 Rutgers U 做助理教授，我還在 AT&T 服務。我問他：目前統計界熱門的研究專題是什麼，他就給我一個名詞：bootstrap。

Bootstrap 的最初論文講的是 iid 的 data。那時我手上有一個抽樣的計劃，但是被抽樣的母體的有限的，在有限母體上最簡單的抽樣都不可能是 iid。我就提出一個問題：在 non-iid 的情形下，如何做 bootstrap。

我已經忘了這問題我們反復討論了多久，但是漸漸地我們對抽樣的目的對於統計學的真正意義：要被弄清楚的是整個母體而不是那幾個特定的參數。所以，從一個有 N 個對象中用 simple random sampling without replacement 取 n 個樣本，我們要估計的是 N 個數值。如果 $N/n=k$ 是一個整數，那麼我們就把這 n 個樣本每一個重複 k 次就好。而這樣的一個新集合，可以證明它就是原先被抽樣的母體的 MLE。

我和他大概有十幾篇還算可以的論文，就不提了。
因為拒收輓聯，我就擬了一幅：

四十年相知，藝出同源，談笑歡飲，不曾忘先師存餘惠；
七八篇合著，血汗滄桑，時有醍醐，再回頭燈火已闌珊。

二、危機與心態（短語）(2025.10.10)

危機就是機會，困難就是學習，大地的創傷我們還是要往前。好句子，也是好心態。

三、生日只能有蛋糕嗎——水調歌頭(2025.10.6)

今天是中秋節，想起一篇舊作，是吾友黃文璋教授還在人間的時候在下作的。特此貼出來來紀念當年對酒當歌的美好時刻。

生日只能有蛋糕嗎

《水調歌頭》 2nd edition, 2020/8/29

文璋注：生日只能有蛋糕嗎¹？愛酒不愧天，當年有幾位頗得酒中趣的小友，包括慶剛、文瀚、宗記、育杰及逸輝，突發奇想，在相互慫恿下，每人買一瓶自己生日的年份酒。可不是要立即舉杯邀明月，預計各自七十歲生日時，才邀諸親友一起享用。這主意太好，陸續有同好加入。只是對年輕人，七十歲，還多麼久啊！慶剛率先開酒了，沒關係，千金散盡還復來，他另提供一瓶交宗記保管。酒香餘韻久久未散，趙老師作的詞。更讓這份回憶雋永長存！期待明年趙老師來南區，帶領大家留其名。

民德也注：文璋，黃姓大教授，與夫人羅夢娜教授，數十年來，號召臺灣統計界英傑，連續n年在臺舉辦“南區統計研討會”，並讓“南區”的定義逐漸成為“基隆以南”，氣魄宏偉。俠骨英風，長歌慷慨，人稱南霸天。三十餘年前，余曾三次南下，希望能招攬此賢夫婦北上，均未能如願。雖然甚憾，但回想起來，未嘗不是好事。因為學術進步，要靠砥礪競爭，各樹山頭，而非一家獨大。所以台灣雖小，歷次統計界精英境外出擊，雖然對手強大，勝負兵家常事耳，但酒筵上從不輸陣，一向精誠合作，勇健以赴。慶剛、文瀚、逸輝、宗記諸君，皆戰功顯赫，不愧鐵血之輩。

人擠店便小，席外吵翻天，
恰逢諸位英傑，藏酒許多年。
不論新醅陳味，況共時憂舊淚，微雨共春寒。
教授幾多位，同在好人間。

酒雖厚，人待醉，夜難眠。
世間萬事，來去唯有地球圓。
儘有前非難改，總是貪嗔癡愛，大衍本不全。
再盡千樽酒，何必訪嬋娟。

次东坡韵，民德戲作

¹ 有點像張清芳唱的：“愛人不能是朋友嗎 …” (Men's talk)

四、虞美人——賀南區 34 屆統計研討會(2025.6.18)

虞美人 双调五十六字，前后段各四句，两仄韵、两平韵——（南唐）李煜的版本，賀南區 34 屆統計研討會。

要將國事管得好
統計最重要
數據都有一些差
沒有常態怎樣分析它

南區妙手年年在
誰說朱顏改
問題到此已難求
不問人智是否皺眉頭

五、根號負一(2025.6.18)

今早讀到老友李家同的一篇貼文：《我的追憶專欄(64)電容器和虛數》大有所感。想在他的FB貼文上留言，但是我想講的又長了些，所以就自己就貼了一篇呼應一下。

即使研究多年的都是不用實驗室的那種工作，但大道至簡的直覺，還是有一點的。下面的，也許已有高手得到某些看法或結果。這不是詩，我只是儘量讓它順口而已。

人類發明多得很
有一個叫根號負一
世上事虛虛實實
它就是真正的虛
我也唸過複變函數
因為曾讀過數學系
但若說真的用在刀口上
卻說不出深刻的道理

我當然也曾用過
但覺得最多只碰到毛皮
可是相關的公式
有些真漂亮無比
大道至簡，你研究的結果如果漂亮
就差不多等於得到真理

在我們面對的世外
好像有另一片天地
它不算太極陰陽
因為無關男女
它也沒啥對稱
我猜相關的也許是兩儀
這樣說來有一點玄
但我的不解本來就是什麼是虛

有一個變換叫做傅麗葉（和羅密歐沒有關係）
應該是奇妙的東西
它可以將一個在A空間難解的問題
換到另一個空間B來處理
有一點傳送陣的味道
用它可痛快地證明中央極限定理
沒有這樣的一招鮮
哪有現代的機率和統計

道家說道法自然
但大自然里看不到根號負一
這後面的道理必然很深、不是水很深
五行八卦風水等等應該全不能比
我猜甚至在另一個位面
根本不包含銀河系
不算冥界，無關輪回
如果你穿越到了那裡
我們的學問極可能只算個屁

是不是時間是倒退的？
是不是能量和質量可以為負，甚至為虛？
也許是完全另外一套的物理

我問了一個物理學家，2020/7/19 李定國說：根號負一是量子力學與古典物理最重要的一個差別之一。所以大自然有它，只是不仔細去看就看不到，它到處都在。傅麗葉變換也與量力有密切的關係。
我的回信：

T. K.，謝謝，我在 Bell Labs 的時候，知道搞通訊的那些人，很重視複變函數，但那時事不關己，也無餘力，便放在一邊。大3時學這門課時，老師也不夠力，講不出味道來。但 complex variable 這是一個完美的系統，例如：

1. n 次多項式，必有 n 個根——這在實數系統里，全做不到。所以，覺得這是寶山，卻無力挖之。
2. analytic function 圍著一條封閉曲線的積分為 0。
3. 歐拉公式 $e^{i\theta} = \cos \theta + i \sin \theta$ ，這樣漂亮的公式。我不知道 $\sin(i)$ 是否能表成 $a+bi$ 的型式。當年問不出這樣的問題，現在沒人可問了。型式上，我會將它用 power series 展開，不管 converge 與否，先猜答案……

時間的本質是甚麼？物理界是如何看的？

六、青年守則第 13 條(2025.6.9)

我小時常學到「xx 為 yy 之本」這樣的教育，今天我想再加上一條：實驗為解惑之本。——這句話還是有一點統計學的意義的：我用了 2.5 小時，發現了家裡的 2 台除濕機其實都還沒有壞。

七、苦難的歲月(2025.5.18)

君子出處，先生不疑。卅年弦歌，盡清華矣。溫柔敦厚，不求不忤。桃李不言，下自成蹊。

山不在高，有仙則靈。系不在老，無情不成。先生獨慈，眾人皆狠。而今而後，火熱水深。

1993 年 7 月 12、13 日，在中山大學舉行的第二屆南區統計研討會，為慶祝姚景星老師自臺大數學系榮退，在第一天(7/12)的會議最後，大會送姚老師一紀念牌。牌上的題詞是我做的唯一的一首四言詩。

我 1957 讀臺大數學系，現在回想起來，那時的教授們好高騖遠，儘開些他們自己都不太懂的數學。有的只靠一本筆記，上課只是抄黑板。有的選一本很難的有名的經典教本，500pp 以上，但常一學期只能教 3、4 章。例如群論，環論，拓撲學，張量分析，微分幾何，泛函分析，近世代數……總之，我大學四年讀得很苦，基本上，連交女朋友的時間都沒有。是數學是那麼難嗎？當然難，要不然，為甚麼 25 個人考進去，最後只有兩個學生四年畢業？那時，大專聯考的錄取率只是在 10%，學生們能上臺大，IQ 和勤奮度，是不必懷疑的。在那時的經濟條件下，每一個學生都背負了家庭的出人頭地的期望。在那樣的條件下，一個系有甚麼權利讓學生七、八年才拿到一個學士學位？我這一班的其他同學都眼明手快：能轉就轉。我下一屆的，有五個學妹很爭氣：她們聯合起來，把沈教授賴以鎮山門的秘笈：高木貞志寫的《高微》找出來，將困擾了同門很多年的護山大陣的陣眼，一一破解。畢業時，她們包辦前五名。

後來的發展，因為當兵、出國等事，我知道的極少。但 70%以上的東西，後來都沒有用過一次。要知道，1968-78，我可是在 Bell Labs 工作，可以說靠的就是數學、統計學上面的知識吃飯。姚老師的我能記得的好處是：他不坑學生。所以，我對他還是很感念的。

八、最長詞牌很難填嗎—鶯啼序(2025.5.16)

最長詞牌很難填嗎

鶯啼序 最長詞牌 (3rd edition)，次“詞如七寶樓台，炫人眼目，拆碎下來，不成片段”的吳文英原韻。此類詞最好改，因為缺乏中心思想，只要加上一點自己的看法，就多了味道。次韻他的詞，不是仰慕，而是因為不必去查相關韻表，就知道絕對不會出韻。(小序作於 2020/6/8) 2020/12/4 又改。

清明過完漸久，愁尋常門戶。
雨未盡、散入離愁，但覺忽已年暮。
煙塵外、平常寄問，抬頭不見青青樹。
有餘情還在，心思但比飛絮。

如此良宵，為誰今夜，立中宵風露？
夜晚矣、瑩燭依然，電郵無需尺素。
問來人、平安是否？還就去、詞填金縷。
且隨君，小渡松江，莫驚鷗鷺。

老來白髮，朝有青絲，人生如逆旅。
撩落去、平生何礙，率性便好；忍又如何，最多苦雨。
長歌如血，悲多無淚，途窮到此應何去？記當時、坎坷無人渡。
都能思量，何需功過平衡，吾輩終成塵土。

荊卿帶劍，易水瀟瀟，盡衣冠白苧。
憑誰問、於期頭去。一擊千年，徐亡無功，破陣有舞。
悠悠古史，平凡而在，薄林秀木隨風去，有芳菲、何必尋樑柱？
傷心不是行人，千古誰評，是非是否？

給老友的信：

漢嬰，
文字遊戲有時很費功夫。我填了一首“最長詞牌”，搞定了平仄韻腳，用了整整一下午。完全是硬湊。以後再不幹了。
你是我唯一可 show off 的朋友。

民德 2020/6/6 人在臺北

附吳夢窗原詞：

殘寒正欺病酒，掩沈香繡戶。
燕來晚、飛入西城，似說春事遲暮。
畫船載、清明過卻，晴煙冉冉吳宮樹。
念羈情游蕩，隨風化為輕絮。

十載西湖，傍柳系馬，趁嬌塵軟露。
溯紅漸、招入仙溪，錦兒偷寄幽素。

倚銀屏、春寬夢窄，斷紅濕、歌紈金縷。
暝堤空，輕把斜陽，總還鷗鷺。

幽蘭旋老，杜若還生，水鄉尚寄旅。
別後訪、六橋無信，事往花萎，瘞玉埋香，幾番風雨。
長波妒盼，遙山羞黛，漁燈分影春江宿，記當時、短楫桃根渡。
青樓仿佛，臨分敗壁題詩，淚墨慘澹塵土。

危亭望極，草色天涯，嘆鬢侵半苧。
暗點檢、離痕歡唾。尚染鮫綃，鞦韆鳳迷歸，破鸞慵舞。
殷勤待寫，書中長恨，藍霞遼海沈過雁，漫相思、彈入哀箏柱。
傷心千里江南，怨曲重招，斷魂在否。

九、七絕二首（忠孝醫院即興）(2025.5.8)

一大早，太太和我去北市忠孝醫院看了好幾個病。最後，在等候叫號的走廊上，看到了叫做“明維心靈空間”的一個有點像畫展那樣地掛了好多幅畫。畫的好壞我是不太懂的，但是有七幅畫的標題都是七個字的好像七言詩句一樣。我就先抽樣再排列組合一下，得到下面的不知是否合律（但至少押韻）的七絕：

雪鵲穿樹喜新晴，
綠葉成蔭子滿枝，
千嬌萬態破朝霞，
花間覓路鳥先知。

落選的三句是：
幽情逸韻落人間，
嫩紅凝脂染春山，
清情高韻傲霜雪，

我還捨不得丟掉，所以就加上一句：
忠孝美女勝貂蟬。

這樣就又是一首不知是否合律（但至少押韻）的七絕了。

十、拜年(2025.9.4)

去看看別人是怎樣拜年的：

馬上要過年了，我現在這裏提前給你拜個年，祝你一帆風順二龍騰飛三羊開泰四季平安五福臨門六六大順七星高照八方來財九九同心十全十美，百事亨通千事吉祥萬事如意！

來源：小說王-台灣最大小說網 > 都市言情《寒門大俗人》 第 116 章。

See: https://www.twking.org/218_218048/107873996.html 作者：畫筆敲敲

十一、江河旋律(2025.7.5)

下面的一段文字，是我的大師兄王鼎鈞的文筆，收在他新出的一本《江河旋律》（這是鼎公的百歲自選集裡面的第一篇中的第一段）：

鄉愁是美學，不是政治學。思鄉不需要獎賞，也用不著和別人競賽。我的鄉愁是浪漫而略近頹廢的，帶著像感冒一樣的溫柔。

廢話就少說了：這一段是我讀過的最好的白話文。

十二、寒門崛起(2025.5.26)

在網上讀到一些妙文，特別選了一段，以饗臉友：

食物體的詩詞

取材自“朱郎才盡”作：《寒門崛起》 第 1528 章

黃錦看到嘉靖帝難得心情大好，機會不容錯過，趁機躬身上前小聲道：「聖上，您今天還沒有進膳呢，為了天下蒼生百姓考慮，您也要保重龍體啊。奴才讓御膳房進些吃食，您多少用一些吧。」

「嗯，方才看到朱平安，朕就禁不住想到他當初做的那首糖醋魚和麻辣翅孰更下飯的詩詞」嘉靖帝微微點了點頭，提到朱平安就禁不住露出了笑意。

「呵呵，聖上說的是，小朱大人做的那首詩，奴才也還記得呢。還有戲改的那什麼'故人西辭黃鶴樓，千里迢迢買魚頭!'、'君問歸期未有期，紅燒茄子油燜雞'、'曾經滄海難為水，魚香肉絲配雞腿』等，小朱大人說這是什麼『食物體』詩詞。這說著，奴才就有點饞了呢，要不讓御膳房都安排上，奴才試菜也能解解饞」。

黃錦附和著笑不攏嘴，倒背如流的將朱平安曾經進獻的食物體詩詞背了幾句。

「嗯，食物體詩詞，呵呵，是挺下飯的，既然黃伴也饞了，那就讓御膳房都安排上吧。」

「嗯，食物體詩詞，呵呵，是挺下飯的，既然黃伴也饞了，那就讓御膳房都安排上吧。」

十三、夢微之(2025.4.27)

近來，老同學們的歡然重聚越來越難了。道理很簡單：我們都相當老了。前一陣，我們所尊敬的一位師母以 92 歲的高齡仙去，在台的同學能去的人只有 5 位。所以，當我讀到白居易的這首詩時，感覺是十分強烈的。

《夢微之》（唐朝詩人元稹（779—831），字微之，與白居易同科及第，並結為終生詩友。）

夜來攜手夢同遊，晨起盈巾淚莫收。
漳浦老身三度病，咸陽宿草八回秋。
君埋泉下泥銷骨，我寄人間雪滿頭。
阿衛韓郎相次去，夜臺茫昧得知不？

十四、中藥也在找搭檔(2025.4.23)

從我的老班長（師大附中高 41 班）那裡轉來，特此貼上，以饗臉友。

《中藥也在找搭檔》

枸杞愛上了菊花，眼睛就亮了。山藥愛上了靈芝，血糖就低了。三七愛上了山楂，血壓就降了。決明愛上了荷葉，大便就通了。蓮子愛上了欠實，脾胃就運化了。黃耆愛上了當歸，氣血就順了。茯苓愛上了薏米，濕氣就祛了。金銀花遇上絞股藍，發炎就消了。蓮子心遇到了麥冬，心火就清了。酸棗仁遇上了龍眼肉，睡眠就好了。您愛上了養生，美麗健康就來了！在「群組冒泡」是預防老年癡呆最好良方。

世界上最好的長壽藥是：喝水、睡覺、走路、唱歌、群組冒泡這五樣，最重要的是，它們都是免費的。美好的一天從別忘了這五種「長壽藥」開始！大自然的秘密 上蒼真的好神奇！

你說巧不巧？人體有 365 個穴位，一年剛好 365 天。人體有四肢，一年有四季。人體有 12 條經絡，一年剛好 12 個月。脊椎有 24 節，一年剛好 24 節氣。人有 7 竅，一個星期有 7 天。人與大自然完全吻合！厲害啊！所以我們一定要遵循大自然的規律！一日之計在於晨，一年之計在於春，祝您幸福快樂，平安健康每一天！讀完後，不要忘記給朋友看了！分享是一種享受，也是積福。

十五、我領堅振(2025.4.19)

前年今夜，是我領堅振的日子。我是 1960 領洗的，在台大對面的聖家堂，1965 結婚的，在 Berkeley 的一家小天主堂，然後一直忙著，到了忘了自己是天主教徒的這件事的程度。然後，隔了 58 年，在一次好友的告別式中，看到十字架上的耶穌，忽然心有所感，覺得天主是在給我一個機會。於是我用 Google 找到離我家最近的天主堂，對神父說：我需要一位能把我拉回天主教的人。

3 次重大聖事，都是我的老同學李家同做代父。

十六、集句(2024.4.27)

無聊時做的集句——意思是：每一句都是抄來的。我就更進一步，連說明也是抄來的：

宋代則有王安石、李綱和文天祥等人的集句詩。

王安石集句詩又稱「荊公體」，〈胡笳十八拍〉為其中巨製，是王安石在「以文為戲」的創作觀念下，嚴謹從事，「用全力出奇」之佳構。具有「創作迅速」、「對偶工巧」、「改易前人詩句」的特色。李綱寫〈胡笳十八拍〉的動機不在替蔡琰增添悲憤之聲，而是澆個人胸中塊壘。就作品的藝術成就而言，李綱〈胡笳十八拍〉的各拍之間沒有明顯的敘事時間架構，全篇意象與典故十分集中，時見輾轉重覆，不像蔡琰流暢陳辭，也不如王安石渾然天成。文天祥〈胡笳曲〉較王安石和李綱的集句詩更具有音樂性，160句全援引杜甫詩。〈胡笳曲〉融合了蔡琰和杜甫詩的格調，將王安石嚴謹的集句模擬、李綱的藉詩言史，融合為兼具忠愛與親情的詩篇。在作品內容方面，王安石客觀描寫蔡琰故事；李綱則全無蔡琰史實；文天祥藉蔡琰遭遇寫個人離亂身世，各有千秋。²

我的集句如下：

齊大非吾耦³

楚客不堪聽⁴

獨釣寒江雪⁵

徒有羨魚情⁶

結論：集句的規矩是很多的。但是，規矩越多，chatGPT就越厲害。所以，以後，做集句的人會越來越少。

² 來自：南宋〈胡笳十八拍〉集句詩之書寫及其歷史意義，by 衣若芬，新加坡 南洋理工大學中文系；刊於《浙江大學學報》42 卷 1 期（2012 年 1 月），頁 128-138。

³ 人各有耦，齊大非吾耦也。」典出《左傳·桓公六年》

⁴ 出自唐代錢起的《省試湘靈鼓瑟》：善鼓雲和瑟，常聞帝子靈。馮夷空自舞，楚客不堪聽。苦調淒金石，清音入杳冥。蒼梧來怨慕，白芷動芳馨。流水傳瀟浦，悲風過洞庭。曲終人不見，江上數峰青。

⁵ 出自唐代詩人柳宗元所作的五言絕句《江雪》

⁶ 唐·孟浩然《望洞庭湖贈張丞相》詩：八月湖水平，涵虛混太清。氣蒸雲夢澤，波撼岳陽城。欲濟無舟楫，端居恥聖明。坐觀垂釣者，徒有羨魚情。

十七、年會(2023)

和統計學有關聯的教授們，頗有幾位能飲且好（第4聲）飲的高手，幾乎每年新春時節，都有一個值得一去的餐會。在下也是能去一定會去，除了今年小恙，就只好遺憾了。我只好給主辦人去了一函，曰：

初四的聚會，我雖不能來，但是還是有點嚮往的。
附上一首七律，祝來聚會的諸君子和淑女，新年快樂。
虎虎生風吹玉兔……

趙先生今年不克參加此一年一度的“年會”，特地惠賜一首詩，頗有古人之風，我在餐會當天拿出手機，朗誦傳來的詩，朗誦畢，獲得滿堂彩，當然不是因為我唸得好，喝彩當然是對詩。附上當天朗誦時的錄影。

十八、敬悼黃文璋教授（2023.11.21）

敬悼黃文璋教授

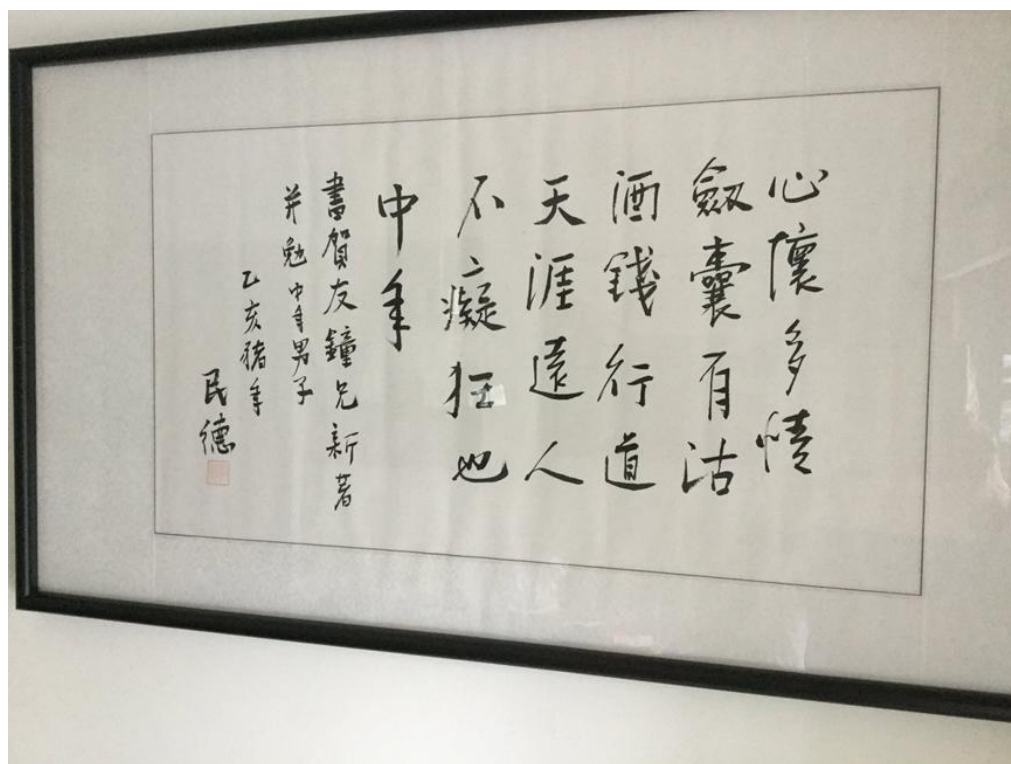
一早就從 Line 中讀到黃文璋教授的不幸消息，十分傷心。這是一位我極為看重和信任的朋友，他對於台灣統計學界的貢獻，知者極多，不需我說。反正他的仙去，是我們心中的極大遺憾。

《若望福音》十二 24：我實實在在告訴你們：一粒麥子如果不落在地裏死了，仍只是一粒；如果死了，纔結出許多子粒來。文璋不止是能結出許多子粒，而且還是結出許多非常優秀的子粒的那種教授。我不如他。

十九、對聯一副(2022.2.1)

無事不登三寶殿
過年才打四方城

二十、賀友鐘兄新著(1995.6.1)



二十一、轉貼兩則

民德兄：

我未讀紅樓後半，所幸論詩一章，確是佳作，將文學批評與生活描寫合在一處，誠少見創舉。批評已同五十年前盛行的 New Criticism，且將逐句品評改為邊寫邊評，又一突破，恐西方未有。唯本章文筆，尚待洗鍊，此或早期臨本小說之通病也。

儼安

文珊路 竹筠時筆
12.7.

數學老師與老婆離婚
臨走女方在離婚書背後
寫著：

$$145 \times 154 \div 4(1+1)80$$

事後，老教授用盡各種
數學專業，仍苦思不得
其解。索性上課時在黑
板要求學生解答，某學生
脫口而出——

一事無成 一無是處 死玉八景

卷二 詩詞卷

一、過沈從文宅(2005.8.8)

邊城人已逝
湘西往事多
瀟湘行腳處
夫子竟如何

在鳳凰城作，湖南

二、某一個晚上的小詩(2006.2.1)

有酒又如何
誰解其中味
天晚正愁余
問我何由醉

詩是在蘇州干將路的貴賓樓上想的
座中的是蘇州大學的幾位老師

三、江湖小隱(2006.6.6)

遠來誰是客
未老莫還鄉
江湖小隱處
煙水兩茫茫

在蘇州的十一樓上，看盡的是江南的煙雨、流水小橋。
仍然是遊人只合江南老。

四、武漢三鎮(2006.8.8)

六十八年墜世間
間關萬里終需還
城廓已非人亦異
朗吟飛過只惘然

這是我出生地，這是第二次來。當中是一片六十八年的空白。
歸來的眼神，是丁令威的嗎？匆匆而過的情懷，是呂岩的嗎？
這次我是從岳陽過來，由武漢去。一個甲子是太久了。

五、蒼狗(2006.9.9)

青衿一別五十秋
蕭蕭白髮漸盈頭
緣有西窗共秉燭
也曾萬里覓封侯

相見時難別甚易
只有歡笑情如舊
應知明日酒醒處
蒼狗白雲兩悠悠

同學會中作，時在日月潭的哲園， 2006/9

六、2006 歲暮雜感(2006.12.12)

世上桃源何處尋
小隱於市止於心
閒來不問天下事
也無憂憤也無情

七、給〈北風〉的歌者(2007.2.13)

北風微冷
在不算傷心的酒店
今夜消魂
明日安歸

我在「上洲」見到他，那是個居酒屋。賣的是酒、歌和寂寞。不記得姓名，乾過幾杯酒，是一家小出版社的老闆。

上洲已然歇業，我也再也不去八條通。
記得的是他愴涼的歌聲，唱的是張鎬哲的〈北風〉：

翻開陳舊的往事 看見一身滄桑
走過陌生的地方 我回到異鄉…
北風又傳來熟悉的聲音
剎那間讓我突然覺得好冷
彷彿在告訴我走得太遠
有沒有忘記最初的相約 ….

八、念奴嬌(2007.9.24)

人遠江湖，依稀是，千里姑蘇歸路。
夾岸青青垂楊柳，怎繫得蘭舟住？
波臥橋虹，湖憑燈影，但明滅如訴：
平林漠漠，不知鄉關何處。

未老，休去填詞，
紅翠溫柔，把青春都負。
黃梅微雨人去後，乍暖輕寒淒楚。
水岸風流，火樹花飛，共波上水舞。
天晚愁余，正覺人間微苦。

2007/6/15 first drafted, 2007/9/8 revised.
Went with neighbors to Golden-chicken Lake just before leaving for Taipei. Lake is well developed, all too well, with all kinds of modern structures, plus trees, waterfront, green turf, music, fountain and lights. It is no longer the type of Suzhou I have been thinking about.

九、過柳永祠(2007.9.24)

九曲溪驚豔後，走過宋街。
宋是醬子的嗎？
古雅的造型，現代的建築，穿越的過客。
——乃不知有漢，無論魏晉。
吆喝著：進來參觀一下啲！

街盡處迎客坊。
說不上驀然相逢的喜悅，
有些意外是真的。
不記得你是閩人哪，
但總覺得曉風殘月楊柳岸該是江南的婉約和瀟灑。

憔悴總是真的，何況在酒後。
衣帶寬是瘦身的結果
（用薰衣草的香精吧）
淺酌低唱的浮名，記得你總是夜店的人氣指標。
六百年，為伊消得。
林間的風，依稀是宋時的一往無悔。

是蠻安靜的，
客少，季淡。
應該的：已不用井取水，
不用你的句子撫傷。
打著你名字的小賣連一本《樂章集》都沒有。

真是寂寞

2007年9月3日，去武夷山竹筏漂流。溪水九曲，風景極美，終矣處捨筏登岸，有街若宋時焉。街盡處完全意外地遇到柳永祠。
柳永詞：她的溫柔清麗與無奈，常陪我度過需要溫柔也有些無奈的夜晚。就有一矣遇到朋友的感觉。久仰了，我說。
該有一个可以敬你一杯的方法吧？可以燒一首詩給你嗎？至少，該有一副紅牙拍板吧。(2007/9)

十、八聲甘州(2007.9.24)

行萬里，幾度舊清秋，且拈花一笑。
問世間閑情，雖長左右，更何時了？
朝來還是青絲，貪歡需早，
漸多有華髮，情人亦老。

誰云乘風可去？但時有清歌，廣陵甚妙。
看月華如練，有素娥曾到。
漸漸如鉤，情深處，腰圍漸少。
登樓看，湖光藹色，煙波浩渺。

2007-9-15 first draft, in Suzhou, Yang-cheng Lake side.
Near mid-fall Festival

十一、子陵釣台(2007.9.24)

不肯做官而已，何為高士；
釣魚豈能養家？不是窮人。

子陵釣台所作聯（2007 sometime in May）

十二、今古四喜詩(2010.3.3)

古：久旱逢甘雨，他鄉遇故知；洞房花燭夜，金榜題名時。
今：錢多事少離家近，位高權重責任輕，睡覺睡到自然醒，數錢數到手抽筋。

作者不詳,但是現代的人和古代的人,想法真是蠻不一樣的

十三、生查子(1964.2.1)

但解別時難
不堪摘紅豆
長日終慙慙
鏡裡朱顏瘦

再見是何期
唯約黃昏後
憔悴儂和我
只有情如舊

十四、秋興(1964.2.1)

西園來秋色，東閣整紅妝，不知相思苦，但覺欲斷腸。
可堪孤館歇，不耐清曉寒，夢裡身是客，只合在鄉關。
日暮江湖遠，修竹倚與誰？榻前雙燭淚，一寸一寸灰。
十年盟滄海，寂寞此一時，不將心中事，說與俗人聽。
孤光豈自賞，肝肺非冰雪，扣舷方一笑，唯我與明月。
天地之悠悠，時令之寂寂，此心之澹澹，此情之戚戚。

十五、舊作(1964.2.1)

寂寞誰可語，楚客不堪驚，為有相思意，略無羨魚情。蓋棺豈論定，等閑負薄倖，悠悠白水下，澹澹游子心。

贈解人，當聞弦歌而知雅意也。

一九六四年十月，秋暮，梧桐墜葉，昔日青青，今猶在否？

十六、地震之後的迴想(1999.2.9)

在這樣溫柔的夜裡，我沒有事。
日子只是一連串的過去，
工作，其實並不是生活的全部
雖然，我也寧可微醉。
日子，只是失望。
日子，還不只 routine
活著，已無責任
只是活著，基本上，沒有夢，也無夢碎。
是太老？
也許，只是心軟
世上原無
捨不去的情劫
誰又把誰遺忘
都是誤會
也都是誤會
會讓人憔悴
哭已無淚
路走到何處
話也不到何處
心，和心沒有交界
山動地搖，我沒有太多
感覺
人生是一場機率，
生生死死
都是 Poisson
分配
酒店關門
其實心早已閉
漠漠然，
大半是疲憊。

1999/9/27 夜作，滿心都是寂寞。

十七、北宜公路十六點二公里(2002.2.7)

四圍皆綠樹
欄外是青山
此中有真味
欲辨已忘言

這是間好吃餐館，有一位可愛的老闆。他十分風雅。
2002 年作，座中有羅小華。
這個店已搬到陽明山

十八、某秋日登山作(2003.2.1)

山高秋老矣
日暮欲何之
江晚問何似
人倦我無辭
今宵多少酒
明日兩行詩
天涯問誰去
不似少年時

1999 年作，失去原稿，後又補作

十九、四不詩(1981)

不道青春喚不回
不信相思竟成灰
不與蒼天問休戚
不到黃河誓不回

1981 作，此詩有英譯。譯者為鍾玲，香港浸會大學講座教授、文學院院長，
2004 年擔任香港浸會大學國際作家工作坊主任，2010 年兼任協理副校長。
2012 年成為香港浸會大學榮休教授。

Never do I believe that the fresh youth can never be recalled
Never have I thought that my longings will be burned to ashes
I would never question Heaven for the cause of my weal and woe
I won't turn back until I reach the Yellow River

卷三 散文卷

一、舊函(2002.2.6)

忽接來函，情懷如舊。上帝造萬物，有一事永遠公平：每人每年都恰好老一歲。時光如昨，弟亦於今年二月退休。仍然每日窮忙，雖不是日理萬機，自知能用心工作的時間不多，總好像是要趕著將事情做完。目前還欠兩本書未寫畢，還在另開戰場，有一點不知今夕何夕。

基本上我已放棄寫文藝作品，不論是詩歌散文小說，因為年紀到了，便自知不是那種材料。所作都是雜事，除政治外都有一點關心，或者，對萬事都視作水月鏡花。

人生隨緣留住，最好心無挂礙。結廬城市，雖見青山但亦不覺嫵媚，而車馬喧也視若無睹，聽若無聞。真到了看山看水仍然是山是水的境界。第四遍又讀黃易的《大唐雙龍傳》（這是好休閒書，凡五百萬言，推薦），夫人病之。我曰：余自總髮以來，一直工作至今，基本上從來沒有休息過。到了這個年紀，妳還盼我上進不成？

目前不會長途旅行。因為一方面不捨到坐頭等，另一方面又覺得到了這把年紀還是擠經濟艙，內心不太平衡也。

台灣的政治，遲早自己會解快，因生命會自尋出路，且天下大事，分久必合，合久必分。是非成敗總成空，包括漁樵江渚。誰還記得西遊記裡一開頭出來過場的魚夫樵子：他們用一首一首的「西江月」，爭辯著你的山青，不如我的水秀？非干病酒，不是悲秋，因為新來未瘦。去年在上海的「東方明珠」那，忽然遇到一群有九千人叫「夕陽美」的老人旅行團。他們從東北的老家一路玩下來，浩蕩得像軍團移防。這樣下山的夕陽，也是蠻燦爛的。

二、委托書的世界(2003.2.1)

自從我們可以出售投票委托書以後，國內的選舉就完全沒有賄選了。一切活動，再也沒有激情，既不需要插旗，也不需要造勢，更不必辦餐會。選舉辦得清清爽爽、乾淨利落。事實上，好些候選人很早就知道他們的票數夠了。他們可以立刻把精力用在正經事情上，比如說政策啦、如何運作修改地目啦、要不要三通啦，要不要再買武器啦，等等等等。

當然，還是有不肯出售委托書的選民，他們可以去聽政見，能用口才和理想說服選民的候選人還是有空間的，只是他們用不著大聲疾呼了。這一類的選民比較要聽真正的道理，要騙他們也比較難。

這個觀念可不是容易得來的，但這麼多年以來，我們抓賄選抓得實在是太辛苦了。後來大家大澈大悟，反向操作，才發現世上的道理還是經濟的道理比較務實，大家也很容易便找到了邏輯基礎：國家如上市公司，選民如股東。如果公司的董監事改選可以用委托書，選舉大小官員和代表委員又有何不可？沒有股東想公司倒閉的，當然也沒有選民想國家出問題。退後一萬步來說，就算將來有一天出了問題，也至少是他們拿到代價之後的選擇，股東認賠嘛！誰要你選錯股？誰要你跟著炒作？

委托書雖然面額只是十元，卻是有價碼的。每個選區，每一天都不一樣，和股票一樣——當然啦，委托書是有價證券，只要找一份當天的晚報就知道最新的行情了，完全可以用電腦下單來進出。它的價格由供需來決定，當然少不了有第四台的專家來每天解盤，讓大家可以選擇最好的買點或賣點（非候選人當然可以設法低買高賣，風險當然是有的，但最糟的也不過是最後將套牢的部分吃下來，在選舉日將它們用掉）。我們是大家都知道股市運作的國家，這些道理一講就明白，完全不需要教育選民；一切運作透明化，完全用不著樁腳，也不必操盤。所有的送禮、招待通通合法化。不過當你知道你那一票值五千元的時候，當然不會只要一個一千元的禮物。反過來說，哪一個傻候選人會用送禮的，他怎會知道選民要彩色鍋還是味精？當然是直接收購委托書要有效得多。

誰說台灣的選舉太多？有了這一套機制，選得再多都不怕，因為大多數選舉都會在安安靜靜中進行，沒有甚麼社會成本。何況我們是資訊大國，設計一套不需現場投票、連開票也全電腦化的系統也不會太難。哪裡需要把記者累得像甚麼似的。說不定憲法會明文規定，任何公職，任期都不得超過三個月。這樣，每過一段時間，大家就多一張可以賣的委托書，豈不美哉。

三、關於蔡文甫的《天生的凡夫俗子》(2004.2.1)

作家的最初願望是作品被登出來。再不怎麼樣的文章（包括有公式圖表的學術論文），一旦印成了鉛字，感覺上就立刻不同了。即使是過了二、三十年，偶然翻起舊作剪報，便依稀回到當日的心情。這和看老相片不同，也許是拍相片只是擷一擷按鈕，缺少的是三更燈火下的掙扎，因此也沒有從毛蟲變成蝴蝶的喜悅。

等到懂得投稿、刊載甚至得過一些獎之後，作家就會比較想做一下編輯了。可以生殺別人的作品，而不是自己的作品被別的編者魚肉——不是說上官婉兒（還是誰）生下來就有一把尺可以量天下之才嗎？這是千古的文人夢——是何等滿足自尊心的事情！自古以來，向來是做文章的在下，看文章的在上。做文章的叫做學生或者門生，看文章錄取你的叫老師或者座師。幾千年的科舉下來，那些情結深深地嵌入天下文士最底層的神經。噢，不要以為學術論文不是這樣，我們也有我們的規矩和傳統，一樣也是有若春蠶，至死不悔。

(more…後續段落已遺失)

四、小言(2004.2.2)

中國人說話，有時只說一半；有時甚至只寫一半。

這個標題，來自陳杰作寫的《大染坊》（這是有二十四集的電視劇，台灣還沒演）。這是一位有志難伸的軍人掛在牆上的字：每一個字都只寫一半：小字是「不」少了一橫；言字是「語」少了右側。

不語就是不說話。不該、不能、不會、不想。這和日本人常畫的三隻猴子相仿：不視、不聽、不言。在職時不語，退休了何必再說？

施耐菴在他為《水滸》作序裡說：

人生三十而未娶，不應更娶；四十而未仕，不應更仕；五十不應在家；六十不應出游。何以言之？用逢其時，事易盡也。

大數學家 J. L. Doob 退休後不再給演講。他說：要我講甚麼，難道講釣魚？

下面的詞，便有些淒清了，但頗切題，尤其是在這樣的夜晚：

虞美人

少年聽雨歌樓上
紅燭昏羅帳
中年聽雨客舟中
江闊雲低斷雁叫西風

而今聽雨僧廬下
鬢已星星也
悲歡離合總無情
一任階前點滴到天明

感謝承德兄特意為我的退休用心。也多謝好幾位朋友的鴻文。有幾篇我已看不太懂了，但我都一一讀過——蠻努力地讀過。下面兩句是我略改了一些：

冷眼看花花不語，亂紅飛過秋千去。

讀了多年的論文，也不知是用了熱眼還是冷眼。但是，「酒店關門，我就走」的道理，我是懂得的。

我會去開一家酒店。

五、仲秋時節好讀詩(2004.2.7)

其實甚麼時候都好讀詩。這是因為詩短，兩三分鐘就讀完了。

讀詩最好的一點是全然沒有目的。這年頭，誰都知道做詩人無利可圖。

讀詩的另外一點好處是：它可以一讀再讀。

秋天當然是好時節。君不見：春天不是讀書天，夏日炎炎正好眠。秋天不知如何說，冬天還可等新年。

近來又買了幾冊有韻的「三百首」型的書。都不貴，但頗值回票價。用它們來消磨好幾個黃昏夜晚，過去讀過的、偶然又新發現的、以前沒感覺的、還有以前有感覺但現在沒感覺的……。

一個人在理工的圈子裡過了四十多年，到了晚年，忽然談起讀詩來，是不是有一點奇怪？當然他一點也不會專業。這也無妨，反正四十年來，我一直是在玩票。並且只在特別有心情或者特別沒心情的時候玩。

兒子在三歲時太太就教他床前明月光。太太其實自己從不讀詩。雖然她也算是詩人余光中的學生，大學的時候選過課，當然余先生記不起來。不過有今年有一次好機會和余老師和師母吃飯，又有一張合影為證，余先生只好認賬。

太太教兒子讀詩的道理，是在有客來時讓小子有所表現。小傢伙懂不懂是完全不在她考慮的範圍的。但那個時候兒子的確能背上一些。我想，大概是太太會的他都會。現在兒子還記得多少則完全弄不清楚了，也許媳婦還在教孫女背詩罷。

想來自己也差不多在那樣的年紀和詩接觸。那時外婆和我們住，她時常唸的是：

少小離家老大回，鄉音無改鬢毛催；兒童相見不相識，笑問客從何處來。

那是一九四幾，抗日戰爭的中期。我們是下江人住在重慶的鄉下，每日看見的是日本轟炸機從頭上飛過。大家都是離鄉背井的。我當然不知道那是詩，當然更不知道甚麼是詩。不過三、四歲時，聽得多自然就記得。直到好些年後開始翻讀唐詩三百首，才忽然見到舊識。我們是在離亂中長大的，鄉愁雖有而不濃，因為離亂慣了。但是再見到它仍然十分驚喜。我的啟蒙詩卻並不曾給我多少震撼，還感受不出那裡面的味道。只是小時記得的事物，一直不曾忘記而已。

直到幾十年後在北京見到我的小舅。在四川時的青年人忽然七十歲，然後這樣的感覺就忽然強烈起來。下面的一首，是在回台北的飛機上寫的，賀知章的味道是有一點的：

少小離鄉老大還，昔日青絲今已斑，相逢豈為踐舊約，世情寧不有心酸？
且喜筋骨腰腳健，不愁關山雲嶺難，歷歷未必重頭數，往事如煙自惘然。

（《寄小舅》1991/4 作）

這詩放在清朝以前是不值錢的，因為會做的人太多了。但是物離鄉貴，人離鄉賤，春秋時代的破陶器都是國寶。現在的人能知道七個字一句大概是詩就不錯了。還說甚麼平仄韻腳？我的小舅是那邊戲劇界的耆宿，當年號稱是「鬼才導演」，是有才情的人。詩寄給他，著實讓他感慨了好一陣呢。

到了台大唸數學系，雖然拼命讀一些讀不懂的數學，但是，自以為國文很好卻是一病。大抵是功課上的壓力愈大，想找出口的心情也愈烈。回想起來，我多半的數學統計以外的雜文雜詩，都是在考試過關的壓力較大的時候寫的。因為別的時候去玩都還來不及，好像就沒有話要說，也不需要說。

但在大學作詩卻是因為選了張敬教授的《詩選》要交作業的緣故。其實我在高中遇到極好的國文老師，詩的平仄是懂一點的。第一次交作業，心想字愈少，愈容易配合格律。四言詩太早了，也沒有甚麼我懂的規矩。所以就湊了一首五言絕句。

昨夜風和雨，平明掃落花，塵灰竟委棄，未許逐天涯。

大學時代的我，豈會有這樣的心情？後來讀到辛棄疾的「為賦新詞強說愁」，啞然失笑。原來大家都會玩類似的遊戲。這篇作文分數是不錯，張老師密密加上紅圈，以示鼓勵。並且還加上言語，說年青時詩不要這樣蕭索。那時我住在台北市溫州街的一條沒有花樹的巷子裡，對門居的洛陽女兒是還在唸再興幼稚園的張燕。是沒有落花可以掃的。但是為了製造一點意境，就說不得了。

這詩裡我較愛的兩個字是「未許」，卻是另有所本的。那來自郎紅浣的《古瑟哀弦》，是台灣最早在《大華晚報》上連載的武俠小說。這一折寫的是某少年俠士A遇到另幾個少年俠士，雖沒有相談甚歡，當下結交，卻都留下深刻印象（當然都是帥哥美女啦）。這幾個少年俠士言談裡常提起他們的大姐梅問。俠士A後來在酒店牆上留下一首五絕，前兩句我實在記不得了，大意是說遇到的幾位都頗有可觀，最後的兩句，說的卻是

……微嘆何所恨，未許問梅花！

這詩當然後來被梅問唸到了，並且馬上來電。如果真的發生，那是以前的事，清朝的事。可是，「未許」這兩個字卻被我學會了。

畢業了之後在台南的炮校當兵。那是專業訓練，在五十歲除役之前，我的專長是1202的「野戰炮兵官」，都是拜在炮校之賜。炮校的舊址現在已是成大的一部分。

有一位比我略大的表姐在成大做助教。這是表姐的平方，因為實在的是她的母親是我母親的表姐。可是我們這幾代人都是在外邊流浪的，所以親戚不多，有

一點親就念著。我知道「度小月」和億載金城都是這位表姐帶的。可是我記得較清楚的，卻是她案前貼著的一首詩（她是國文系的）：

前舟已渺渺，欲渡誰相待。秋山起暮鐘，楚雨連滄海。風波離思滿，宿昔容鬢改。獨鳥下東南，廣陵何處在？

韋應物：《淮上即事寄廣陵親故》

那時節，她的膩友正在國外打拼，這首詩不止是明示著她是名花有主，並且思念的情懷，又是淡淡的，又是深深的。後來她住在普林斯頓，和我家距五十哩，時相往還。前一陣表姨過世，再見到她時大家都是「蕭疏鬢已斑」。台北來去匆匆，只是揮手而別。

我愛的韋詩其實是下面的那首：

江漢曾為客，相逢每醉還，浮雲一別後，流水十年間；歡笑情如舊，蕭疏鬢已斑。何因不歸去，淮上對秋山。

韋應物《淮上喜會梁川故人》

雖然是「喜會梁川故人」，但淡淡的哀愁，卻就那樣飄起來了。何因不歸？呵，現代的人何嘗還有家鄉？誰不是飛來，又飛去？

至於為甚麼會愛，它有甚麼好？都不是我們該深究的。讀詩，有若品酒。品酒者只問氣氛對不對，人對不對。酒是如何造的，干卿底事？

詩是甚麼？誰知道？想來也沒法子給一個定義。也許是 $5 \times 4 = 20$ 個字，或者 $5 \times 8 = 40$ ， $7 \times 4 = 28$ ， $7 \times 8 = 56$ 等規格。它也許是一種味道，有時顯，有時不顯。多半的時候打動不了我們，多半的時候也不討厭，除了那些特別酸的。但是，有一天時候對了，場景對了，而忽然那些字，那些詞，那些句子就跳出來。不是撥動心弦，而是熨貼你的感覺。

年青時讀詩有兩個來源。第一是國文課本，多多少少會選幾首詩啦詞啦的。再一個是自己去找。但是讀歸讀，背歸背，偶然也會覺得某一句好，偶然也學父親在自己覺得好的句子上劃上紅圈。但是感覺不強，因為目的還有點世俗：詩記得多了，作文時句子會好看一點。那個時代，考大學作文佔的比重還蠻高。

讀詩的好處是你可以懂也可以不懂。不懂也可以照唸，照樣顯得有一點學問和一點風雅。唸得長了，感覺就出來了，也許是年紀到了，感覺才出來。「人生卻待中年後，爐火是看純青時」，我偶然在某處讀到，大有所感。那時候已過中年。爐火純青談何容易，但是「過中年」並不那樣難呀。

（這是某一年在高雄的中山大學圖書館講稿的一部分。）

六、年輪(2006.2.7)

這是《飄著細雪的下午》的自序

出這本小冊其實想了很久，以前的理由是：現在我再也寫不出這樣的東西了；比較真實的理由是不忍棄之，雖然別人看起來也許覺得是雞肋；更進一步的理此由則是其實自己還是蠻喜歡這些文章的；至於問到為甚麼真的有些喜歡，問題就比較認真了：這裡記錄的，是我成長的痕跡。

這是舊相片留不下的痕跡，因為它是用心靈，而不是用快門去刻劃的。每讀一次，心裡就多多少少地回到那些燈下的夜晚。虛構的情節和半真半假的故事交雜，反應的是心頭的念念。是了，這都是舊夢，它就好在不夠成熟呵。要讀成熟的作品，圖書館裡多得很哪。

青青澀澀的記憶，青青澀澀的年代，用青青澀澀的愛情觀表現出來。在有了責任之後的日子，就全然不是這樣的了。再回首恍然如夢：「曾經在幽幽暗暗反反覆覆中追問，才知道平平淡淡從從容容才是真」。以前的長夜、歡笑、憂愁、傷心、無奈 … 像年輪一樣的，刻劃著那些記憶和非記憶。那是六十年代，男孩女孩還蠻遵守禮義廉恥的遊戲規則，我多半的朋友最後都是和他們的女朋友結婚（甚至有兩個女友的都很少），並且都蠻用心的地經營家庭，這些人多多少少後來都有一些成就，最起碼都是不討厭的自了漢。

時間真是太久了。樹的外皮當然都是風霜和艱困，裡面的心還是溫柔的。這樣的夜裡，溫溫軟軟的回憶，辛辛酸酸的日子 … 人生本來就是這樣的。

七、一壺濁酒喜相逢(2007.2.3 轉貼)

滾滾長江東逝水，浪花淘盡英雄。
是非成敗轉頭空，青山依舊在，幾度夕陽紅。
白髮漁翁江渚上，慣看秋月春風。
一壺濁酒喜相逢，古今多少事，都付笑談中。

這首詞是明朝楊慎所作，清初毛宗崗拿來做為三國演義之卷頭詞。打開三國演義，在第一回“話說天下大勢，分久必合，合久必分”之前，會先讀到這首詞。

內人常不解為什麼我們每逢聚餐都要喝酒，究竟喝酒有什麼好？

其實我們倒沒有每次吃飯都喝酒。而通常就是好友相聚時，因快樂而喝。為自己快樂，也為別人快樂。

我們自己一個人是不喝酒的。平常我們也都是努力地工作，不像蘇軾(水調歌頌)：

明月幾時有，把酒問青天。

我們想問題時就認真地想，不用喝酒。我們不是得意或失意時才喝：

人生得意須盡歡，莫使金樽空對月(李白將進酒)，
落魄江湖載酒行，楚腰纖細掌中輕(杜牧遣懷)。

也不是憂愁時才喝：

酒入愁腸，化作相思淚(范仲淹，蘇幕遮)，
水調數聲持酒聽，午醉醒來愁未醒(張仙，天仙子)。

辛棄疾的西江月差不多可以形容我們喝酒的情景：

醉裡且貪歡笑，要愁那得功夫？
近來始覺古人書，信著全無是處。
昨夜松邊醉倒，問松我醉何如？
只疑松動要來扶，以手推松曰去。

這首詞可說是我輩喝酒的寫照。不是藉酒澆愁(要愁那得功夫)，而且喝了後還要適當地維持尊嚴(不讓松扶)，有時喝完還有一些正事要做的。我們不划拳不行酒令，我們是快樂地喝，喝完最多最多去 KTV 唱個歌，然後安全地回家。第二天按時起床照常工作。不像柳永(雨霖鈴)

今宵酒醒何處？楊柳岸，曉風殘月。

在喝酒時我們奉行二原則：

1. 做人可以隨意，喝酒不能隨意，
2. 敬酒對象要明確。

因此要嘛不喝，要嘛就乾杯。而且也沒有拿著一個杯子由右揮到左說“我敬各位”的。看你要敬誰，才跟你喝。對那些堅稱不會喝酒的人，我們會告訴他：喝酒有兩個臨界點，過了第一個開始舒服，過了第二個便醉了。

因此不喜歡喝酒的人可能從來就未過第一個臨界點。但第二個臨界點是絕對不要超過的。至於第二個臨界點在哪裡呢？那就要靠不斷嘗試，以找出它正確的位置。結論是還是要喝，要繼續喝。

如果沒有喝酒，一桌人通常是正襟危坐，彬彬有禮，彼此保持適當距離。看看這道上的是什麼菜，一個至一個半鐘頭，該上的菜都上了，大約便可打道回府。但要是有了酒，就完全不一樣。通常第三道以後的菜都不重要，因沒有人在乎，大家都專注於找對象敬酒。你只聽過把酒言歡，可沒聽過把菜言歡。一群人忙著喝酒，氣氛融洽，一頓飯吃三個鐘頭還依依不捨。

所以若聽到飯局是在自助餐廳，大家興致便不高。第一自助餐廳通常沒有酒，第二自助餐廳強調的是食物的豐盛，任你吃到飽。每個人起起坐坐，一盤又一盤，恨不得把所有食物都端過來。當你吃冰淇淋時，會有人告訴你，這一球在店裡賣多少錢，你吃了幾球便划得來了。就算餐廳有提供酒，在那種環境下，怎麼喝呢？我們喝酒時，可不是在乎多喝一點才划算。

喝酒時由於精神鬆懈，失去戒心，不要說酒後吐真言，可說是原形畢露。一個人的真性情，在喝了酒便可完全看出。我們不能說只有喝酒才交得到朋友，但喝酒後對朋友的了解會更深一層。我們有一個好朋友，他的網站就叫“趙家酒店”，自稱酒店主人。酒店分為四個部分，第一部分取名為“輕飲小酌之所”。

你以為我們酒量都很好嗎？非也，我們只是有酒膽，我們只是喝酒時不窩囊。我們必須拼命勸酒，光自己喝可是會醉倒。我們離喬峰還差遠了：

喬峰一口氣不停可喝二十來斤，喝完肚子微微脹起，臉色黑黝黝的一如平時，毫無酒意。

蘇軾(臨江仙)說：

長恨此身非我有，何時忘卻營營？

我們的答案是：

一壺濁酒喜相逢，古今多少事，都付笑談中。

不分地位高下，不分尊卑長幼，不分學問好壞，不分海峽兩岸，在酒店裡大家平等，喝了酒便是朋友。

（這是吾友黃文璋的文章，他的作品集是《心在南方》。）

八、寫書是一種情結(2007.2.3)

讓我最有印象的是史努比 (Snoopy) 一再在狗屋頂上打字寫書的情景：

It was a stormy night

是的，寫書是一種每個文化人都不能忘懷的情結。

我也有這樣的情結。但我沒有記日記的習慣，記性又不夠好，又喜歡丟東西：認為以後沒有用的資料，總是先丟再說。因此像永良兄這種自傳式的書，我是寫不出來的。

我們都生在「男兒立志出鄉關」的時代。背景相差不多，不外乎「來來來來台大，去去去去美國」。難得的是和永良兄還算是同行，因此這本書讀起來的感覺就相當強烈了：有太多的東西我也頗為知道（雖然，記不清楚了），有了這樣明確的記錄，也許比時下的「口敘歷史」，可信度還要高哪。

永良兄是一位極好的學者。學問、學風和品味都是一流。這本書講的是一位學者的歷程。簡單地說，就是一步一腳印地穩穩地一路走來，既不求聞達於諸侯，又不浮浮誇誇地嘩取眾寵，當然就有可能做出一些好東西。所謂好，指的是這世上只有少少的幾個人能夠擊節稱許的寂寞。—— 走到高處，豈止是向來蕭瑟而已。

這是為董永良的…《回首學算路 —— 一個旅美學者學算的故事》
（商務出版社）寫的序

九、羞愧與感恩(2007.2.9)

我從小就沒有自以為了不起的想法，這是因為我曾經看過很多厲害的人物，功課比我好的多的是，體育比我好的更是一把抓。我不但不會驕傲，反而總覺得自己不如人家。也因為如此，我一直沒有什麼自以為幸福的感覺。

這種不覺得自己有多少幸福的時代，最後終於結束了，當我到達了印度加爾各達的垂死之家，看見了那麼多赤貧如洗的人，我的想法完全改變了，我從此以後會一再地提醒自己有多麼幸福。我永遠忘不了的是一位十來歲的年輕乞丐，他常握住我的手不放，每次我離開垂死之家，都會回頭看他一眼，他也會揮手和我道別。我因此開始有了羞愧之情，因為我知道我會回到我舒適的家去，而他呢？他如能活著離開垂死之家，也只能回到他當初求乞的地點。我雖然年紀不小，但我仍有事業和前途可談。而他呢？他只有十幾歲，他已命中注定沒有前途了。

自從這次經驗以後，我開始有了完全不同的想法，我發現我之所以能夠在社會上稍有成就，完全是運氣好的緣故。如果我生長在非洲，或者印度非常貧困的家庭裡，我會有今天嗎？如果我生下來就是乞丐，不要說唸書了，我連吃飯都會有問題，極有可能的是我現在早已不在人世了。

自從去了一次垂死之家，回台灣以後，才發現台灣也有不少不幸的人。首先令我感到吃驚的是我國的失功能家庭非常之多，有的是父親不見了，有的是母親不見了，有的是爸爸長年被關在監獄裡，當然有更多的是爸爸酗酒，或者父親一直失業，以致於情緒很不穩定。在這種家庭生長的孩子，能有什麼前途呢？

我的父母一直給我一個溫暖的家，使我有一個可以唸書的環境，也使我遠離世界上很多的罪惡，可是我一直不覺得這些有什麼了不起，等到我看到了這麼多不幸的小孩子，我才恍然大悟，知道自己多麼地幸福。

在我們的社會裡，也有不少家庭並非是不好的家庭，但是非常窮困，在我們競爭如此激烈的情況之下，貧困的小孩子是很可憐的，他們如果功課有不懂的地方，回家往往無人可問，他們的父母不能送他們去補習，不能替他們請家教，他們的功課怎麼能好起來呢？這些窮困的小孩子往往沒有錢買書看，和其他同學比起來，他們的文化刺激少得可憐。

我最近常常慶幸自己不是生長在如此惡劣的環境之中，因為我知道如果我有如此惡劣的生長環境，我絕對不可能有今天的成就的。

也就是因為我自己知道有多麼的幸福，我最近常有羞愧之心。我總覺得我生活得如此之好，實在不是因為我很努力，而是因為我得天獨厚，家庭好，夠聰明，從小就不知道真正貧困的滋味，有這麼多好條件，當然可以在社會上立足了。

羞愧之情使我常有「這可能是我」的想法，夜深人靜，我常想起在加爾各達垂

死之家的那位年輕乞丐，而且也會想「我也可能是乞丐的」，看到在戰爭中喪生的無辜人民，每次看到窮困的小孩子，我也會想到「他就是我」，我就是這種心情之下，寫出「我的故事」的。「我的故事」裡的主角，經常提醒自己可能是一個不幸的人，這種不斷的提醒使他比較有慈悲心。大多數毫無慈悲心的人往往是對於周遭不幸的人毫不知情。有時我們不能責怪他們，在建中唸書的孩子，什麼題目都難不倒他們，他們又如何能想像有人不會解一元一次方程式呢？在正常家庭中長大的孩子們，也不可能知道有些孩子等於沒有爸爸媽媽，因為爸爸經常被關在監獄裡面，媽媽又已經離家出走了，收入豐富的電子工程師們當然也不知道有人經常找不到工作，既使找不到工作，收入也少的可憐。

我是學電腦科技的人，我們常講「虛擬實境」，有一部叫做「駭客任務」的電影，甚至形容人可以生活在一個虛擬的情境之中而毫不知覺，我當然知道這是不可能的，如果可能的話，我們實在應該使我們知道何謂貧困，何謂無家可歸，何謂在戰亂中顛沛流離。雖然這不容易，我們卻應該努力地想像貧困，無家可歸等等人類的苦難，唯有如此，我們才會有悲天憫人的情懷。

羞愧之情也使我越來越有感恩之情，不是嗎？我何德何能，能在社會上立足？如果不是社會給了我一個好的生長環境，我一定一事無成的。

每天早上起床，就可以看到報紙，我當然應該感謝送報紙的人，他們必須好早就起床，才能使我們這些大爺們能夠在六點鐘時刻看到報紙。到了學校，發現大樓走廊裡地磚一塵不染，我知道這是因為一大早，有一位中年婦人每天將地磚拖洗一遍，她看到我一定會熱切地和我打招呼，有時，廁所裡沒有擦手紙或者洗手的乳液用完了，她一面補充，一面還表示歉意。我真的對她心存感激，因為我可以感受到她帶給我的溫暖。作為一位教授，我常常想，還好有麼多已經作古的人做了那麼多的研究，才使我們這些後知後覺者有精彩的書可教，如果我有一些研究的成果，應該感謝同事們的切磋和研究生的努力，我能不感謝他們嗎？

我最近偶然間看到一篇愛因斯坦寫的短文，文章中，愛因斯坦說他一直心懷愧疚，因為他的一生，得到別人的幫助實在太多，所謂別人，他特別強調包含已經死去的人，他又說，他覺得他向社會借了很多的債，因此他說他必須提醒自己，要為別人而活，這篇文章的題目就是「人人都應該為別人而活。」。

愛因斯坦在學術上的成就，幾乎到了前無古人，後無來者的地步，但是他仍然強調他應該感謝別人，因為他的成就也是建築在別人的努力之上。愛因斯坦這樣偉大的科學家，都一再地強調他欠社會太多，我這種人當然更應該有同樣想法了。愛因斯坦說他要努力回饋社會，我更該回饋社會了。

但是如何回饋呢？我有我的作法，我抓一些德蘭中心的小鬼，來做為學生，教他們英文和作文，這些小鬼雖然老大不願意，但是因為我實在教得太好了，他們也就「勉於接受」了我的教導。我們常常聽說弱勢團體的小孩子功課不好，其實都是因為他們沒有人做他們家教而已，一旦有人做家教，成天做習題，背英文單字，數學和英文就不會太差。我也是運氣好，有一位富翁肯捐錢成立博幼基金會，專門幫助弱勢孩子，現在聯電已經開始同樣工作了，而且鴻海也要

在近日內開始。這些事情是我寫「大庇天下阿強俱歡顏」靈感的來源。

可是教弱勢孩子也絕對是一條單行道，我從不覺得我是在做一件偉大的事，反過來我一直覺得這些調皮小鬼們的確帶给了我好多好多的歡樂和心靈上的平安。每次去教書，我就精神百倍，心情極好。（我的那批學生當時心情一定非常沮喪）因為我覺得我自己在做一件很有意義的事，也是使我極有成就感。我常想，我應該感謝他們，是他們使我的人生過得很有意義，「是我應該謝謝你」，就是這樣寫出來的。

講到人生，我有一個奇怪的經驗，我有時被人請去演講，都有聽眾問我一些大問題，比方說：「人生的目標是什麼？」，「人生的意義是什麼？」，我又不是哲學家，學的是電機，如何能回答這個問題？可是我常常問，誰的人生最有意義？每次想到這個問題，我就立刻想到德蘭中心那些修女們，她們的一生，都貢獻給了需要照顧的孩子身上，當她們去世的時候，她們一定可以豪無畏懼的去見天主，對我來講，這種人生是最有意義的。

有一次，我在看小朋友表演舞蹈，一位年青的修女也和她們一起跳舞，而且跳得非常好。當時我在想，這位修女可以像一般人一樣地工作，結婚，有一個溫暖幸福的家庭，但是她沒有走這條路，她將她的一生貢獻給了需要幫助的小孩子。哪一個人人生比較有意義呢？我想任何一個人都會說她選擇的人生比一般人的生有意義。因此我得了一個結論，一個人成天研究人生的意義是不可能得到什麼答案的，重要的是，你要過有意義的生活，這就是「瑪利修女」的來源。

我們基督徒常唱的一首聖歌是：「我有平安如江河」，我們也成天講心靈上的平安，天主教彌撒中更是一再強調平安的重要性，神父最後的祝福也是「主的平安與你同在」，可是奇怪的很，真正有心靈上的平安的人並不多的。我常去德蘭中心，我發現那些好心的修女們永遠有心靈上的平安。我想起德蕾莎修女的故事，有一位富有的牙醫來找德蕾莎修女，他說他什麼都有，就是沒有心靈上的平安，但他感覺到德蕾莎修女有，他說他願意學習她，因為他有的是錢，他說他想隨時跟著修女，如果修女搭乘飛機，他也會買一張機票同行，用這種辦法，他相信他可以領悟到如何能夠得到心靈上的平安。德蕾莎修女聽了他的想法以後，告訴他，既然他有買飛機票的錢，就不妨將錢捐給窮人，既然有時間坐飛機，也就不妨抽空去幫助不幸的人，這樣做，她保證那位富有的牙醫會得到平安的。

我因此寫了「窮人的遺囑」，窮人怎麼會寫遺囑呢？故事裡的窮人，還像煞有介事地在遺囑中註明他將所有的遺產送給某人。幾十年後，我們才知道，他的財產是心靈上的平安，但是心靈上的平安並非白白得到的，你必須去幫助別人，才能得到這種平安。也只有經由幫助別人，我們才能得到平安。

我們做老師的，常會照顧一些清寒的的特殊學生，其實那些非常富有的學生，也是特殊學生，我寫的「特殊學生」，就是大富翁的孩子，他也有他的煩惱，他甚至不敢邀請同學到他家玩。

對於我所有的學生，我永遠告誡他們要有一技隨身，「麵包大師傅」就是這種樣寫出來的，我的意思是說，有的人也許可以走學術路線，唸大學，唸碩士班，唸博士班，但並非人人必須如此，如果你會做麵包，一樣可以在社會上立足。這篇文章一出來，不知道有多少人在問那家麵包店在哪裡。

有一次，我在路上遇到一位老太太，她認出了我，她說她們在幫助一批弱勢的孩子，自從看了我的文章以後，她們就和台北的大飯店聯絡，給幾位孩子做學徒的機會，這些孩子學得很好，也常常送麵包給她們吃。

我的確有一位會做麵包的學生，也常送我麵包吃，「你不要趕我走」，幾乎是真實的故事，有一位普渡大學念電機的學生告訴我一個故事；我的一位高足，楊昌彪教授又告訴我另一個故事，我拼拼湊湊，寫出了「你不要趕我走」，人類的貧困真是人類的羞恥。有時看到旅遊頻道，發現一些奢侈豪華的旅館有給狗吃的菜單。在聖誕夜，BBC 新聞卻報導一則新聞，聯合國被迫將他們在非洲桑比亞的難民救助計畫減半，因為他們沒有錢了，他們只需要八百萬美元，世人真的無法湊到這筆錢嗎？美國人一年在寵物上所花的錢多達 35 億美元，在這個世界上，做一條富人養的狗，真的比很多窮人要幸福的多。

我如何會想寫「阿傑的姑媽」？故事裡的阿傑，應該是我的寫照，我從小就糊塗到極點，我在美國認識了我的太太，談婚嫁以後，必須寫信給未來的岳父大人，我鼓足了勇氣，寫了一封信回台灣，竟然將岳父大人的姓寫錯了。幸好岳父大人認為糊塗的人大概不會是壞人，也就批准了我的求婚。我當然不肯承認自己的糊塗，好在我有一位糊塗的研究生，我就完全賴到他的頭上，他好像無所謂，因為文章中說糊塗的人是有福的。

糊塗的人是否真的有福？我至少知道人不該太精明，有的人事事算計得一清二楚，反而常常發現自己吃了虧，糊塗的人反正不太在乎得失，而且比較慷慨，他們反而往往是快樂的人。

我已快到七十歲了，回顧我的一生，最快樂的事仍是因為我是一位老師，我知道很多老師大嘆現代師生關係的冷淡，可是我從來就沒有這種感覺，我一直覺得我的學生好得不得了，我因此寫了「再給我三十年」，以表達我對我學生的感激之情。我這個人好為人師，那個學生做了我的學生，我就鬧他一輩子，很多學生早已畢業，卻仍被我壓迫每週要做英文作文，而且還要讓我知道他增加了多少生字。這些同學聽說我還想再做三十年的老師，無不暗暗心中叫苦，有一位甚至告訴別人，他後悔當年選了我作為指導教授。

我當然一直有一些令我擔心的學生，當年他們在小學的時候，我曾經教過他們，後來就失去聯絡了，我只知道他們都是中輟生，也沒有什麼競爭力，「第二十一頁」就是在這種心情下寫出來的。

人老了，就會「其言也善」，我真的覺得我是一個幸運的人，而且我越來越知道那些世界上多的是不幸的人，我有吃有穿，就應該幫助那些衣食成問題的人。我當年唸書沒有問題，就應該幫助那些讀書不好的孩子。我自己沒有生病，就應該去探訪長期住院的病患，我自己有一個溫暖的家庭，就應該去看望那些孤

獨的老人。

大概只有像我這種老頭子，會有羞愧與感恩之情，年輕人房貸無著落，汽車買不起，小寶寶的奶粉錢和尿布錢又貴得驚人，不會和我有同樣心情的，但我仍希望大家不要老是只想到了自己，既使年輕人，也不妨想想看，自己是不是一個幸運的人。讀我書的人，如果你願意張大眼睛看一下的話，一定會發現自己絕對是一個幸運的人，世界上比你不幸的人，比比皆是也。

關於李家同

這個人是我的同學：高中，大學，甚至在加州；

他甚至是我的儕相；

但是，以前，我是不知道他會成為名作家的。

關鍵不是他的文筆：我們是同一個國文老師教出來的；關鍵是他的生活態度。

十、Old Letter(2001.12.12)

Wednesday, December 12th, 2001

昨天去買了幾本余秋雨的書。基本上是「文化苦旅」上的舊作，再加上很多精美的照片。於是煙雨江南、西風冀北，大漠飛沙、南海波濤都隨著他的筆意，帶起一陣陣中國情懷。今年是去中國好幾次的一年，南到廣州，只是純為工作沒有去玩；西到長安，去過幼時曾去過的碑林，沒去過的兵馬俑，沒去過的許多「陵」，沒去過的華山，沒去過的法門寺，在這裡我們蠻快樂地渡過好幾天；然後我們去成都：工部草堂，武侯祠，都江堰，樂山佛。都是沾著國內統計界同仁的光，雖給了幾個演講，也都是大家給臉。反而是在北京時間少了，舊遊之地，但我們仍然去看了胡同，坐船從二十世紀壇到頤和園。香港，田昕就實在走不動了。

我們在國內有工作時光，也有快樂時光：西北狼，劍南春，五糧液……，最難忘的是，在峨眉金頂，又是風又是雲，又是高又是冷。還記得那天陳希老分我一瓶二兩裝的二鍋頭，在那樣的日子，二兩是不夠醉的，但是微微的醺醺然，感到的是中國文人的那清楚的一脈從幾千年傳下來的情懷。飲來原不需面紅耳赤，即便是只談著統計，在那樣的人、地、和時光裡，都帶著詩意呢。

只是太太下了峨眉之後血壓便高了起來（並全靠賀家的靈藥壓下來），所以九寨溝就一直沒能去成，這事經過統計界的快速通報系統，恐怕是全國都聽到了罷。

只要一出門，她的胃口就會好。這和不用自己燒飯是絕對相關的。我們也去過波士頓，剛好在九月九日離開那個城，居然是 United Airline 。

那是個一路都頗不順利的旅行。田昕和我都各摔了一交，又扭傷了背，在美國正碰到九一一，回到台灣又正遇上兩百年難見的大水淹台北。總算平安到家，水和電一直停了好幾天。幸運的是，車子雖停在地下室，卻是差了十公分，沒有淹到水。

下午，兩個年青人到辦公室來拉存款和保險。我告訴他們我的年齡，他們很快就走了（你怎麼知道自己過了六十歲？答案是：沒有人來拉你的保險的時候）。天氣又冷又熱，時晴時雨。好像有很多東西可以寫，又好像沒有甚麼到寫的。今年的最大成就，是在此間的《聯合副刊》上登了一篇小說。以後大概是不會再寫的了。自己和朋友都有些吃驚。

十一、義大利心影(1995.10.10)

聽說我們要去義大利，朋友們就熱心地提供資訊：「小心扒手」，「小心吉普賽人」，甚至還有朋友警告：「小心迷香」；——傳言是：有一個團有一晚全團人都離奇地沉睡，然後有人進入房間，拿走現金，旅行支票 …… 義大利是靴形的，笑話的結尾是：「我怎麼在義大利看不出她是靴形的」？

那兒的鞋可是世界一流的，不知和半島的地形有甚麼關係？然兒那些峭壁懸崖的海岸，千鑲萬刻的教堂，懶懶散散但有時又極其認真的風情，斜斜的塔，依稀是血腥和吶喊但已經是夕陽殘照的鬥獸場，被火山灰埋沒而凝結的千古一瞬的名城，既讓人讚嘆，同時又讓人懷疑：這是出墨索里尼的國家嗎？這是出黑手黨的國家嗎？

飛機回到台北的時候，昕和我已精疲力盡。台北的陽光如熾，又亂、又髒、又吵、但充滿了活力，好像每個人都急急忙忙的想完成甚麼事。這真是截然不同的世界。台北也許是二千年前的羅馬，拚命想在建設，但出發點又好像有一點不一樣。但至少，台北人開車和古羅馬人駕戰車的作風，是幾乎一樣的。

西風殘照的，不是漢家陵闕。台北沒有古蹟，台北只有霓虹裝點的繁華，是標標準準的十丈軟紅。台北滿街看到的，都是奢侈，但同時又那麼貧乏。這裡的最大主成分，佔全部 R-平方的百分之九十，只有一個既市俗又實際的指標：經濟活動。所有別的元次，都在它之下苦苦呻吟。這裡不出托斯卡尼尼，不出普西尼；濟慈不來，拜倫也不來。所有的建築都不是為千古留傳而建的。我們建設，是因為大工程意味著大餅。因此，最好是既貴、又不持久。沒有藝術家肯花二十七年，只為要塑一個教堂的銅門。這裡用水泥灌在模子裡造佛像，一次做好多，並且一字排開站崗。規模是不錯的，缺少的只是敬意。

米蘭，匆匆而過

自航空公司簡寫成 LIN 的那一個機場下來，七十三號公車，一張票就帶你到米蘭的鬧區。有些是石版街道，有些是瀝青的。但我們總是找不到觀光指南上說的那一個遊客資訊中心。我們有八個小時，想參加一個三小時左右的城內觀光導遊團。

城市算是出奇的乾淨，公車才停下來，壯麗的歌德式大教堂就挾著讓你透不過氣的美，迎面而來。等我們到了滿是閒人遊客及鴿子的大廣場，我們有看不完的精雕細鑿的歌德式尖塔頂，如劍山石林般的，那樣毫不在乎地浪費著如許的素靜和豪華。

這是義大利最壯麗的歌德式大教堂。自 1386 造到 1813，前後 428 年！馬克吐溫形容為「用大理石寫的詩」。

我們在廣場邊的露天座上吃三文治，喝礦泉水。鋪著有顏色桌布的小桌子，摺椅，以及帆布頂蓬，形成了特有的歐洲觀光情調。用中文讀一首以義大利的大

理石寫的聖詩是感覺不太深的，教堂不是中國文化。感覺深的，而這種感覺隨著在義大利的時間增加而增濃，是好像我們的祖先，除了苦難，沒有給中國人留下甚麼可以看得到摸得到的驕傲。中國人是沒有甚麼祖先遺產的民族。我們也好像沒有想為子孫留下重要的遺產。

匆匆離開。我們只有機會看一下斯卡拉歌劇院。這是維爾地的 Otello 及 Falstaff，還有普西尼的蝴蝶夫人首演的地方。我們沒看成「最後的晚餐」。計程車是找到了那裡，那裡恰巧又不開門。計程車司機也絕口不提今天不開門的事。雖然，兩個外國觀光客叫車去以這幅畫的收藏而聞名的聖塔瑪麗亞教堂，應該不是去做禮拜。我們第一次領教了義大利的計程車：跳錶極快，不論是走的時候，還是等紅燈的時候；單行道極多，很短的路，車反而要走一大節；還有，起跳的那一段錢，幾乎沒有走甚麼距離；最後到了的時候，司機不知碰那一個按鈕，錶上忽然又加不少錢！我們開始懷念台北的計程車了，台北的計程方法，要實惠得多。

我們再乘七十三號公車回機場。附告一點，公車票要在地下鐵入口的小書報店才買得到。有些本地人都不知道，遑論初來的觀光客！好在听從來都不恥下問，我們順利乘公車回到機場。再附告一點，這似乎是我們入境義大利的機場，但我們沒有經過海關。這個機場有國際班機，但沒有海關，異哉。

翡冷翠素描

到翡冷翠住進旅館已是午夜。米蘭到翡冷翠是一家叫做 Avianova 的名不見經傳的小航空公司，雙引擎螺旋槳，虧得台北的鴻霖旅行社的金經理找得到這個小公司！翡冷翠的第一印象是機場狹小，而且和歐洲的那些六點一過，便人人乖乖回家的大城不同，到了半夜，還有閒人在街。這也算是歐洲的特例了，但這和台北的東區，龍山寺或士林的夜市是沒有比的。用台北的眼光看，夜市的閒人只能算小貓兩三隻。

一大早出旅館門便又受到告誡，這次是旅館的工作人員。小心吉普賽人，他們會用搶的！遇到香港的林教授，他說他親眼看過一起，後來中央大學的張教授被搶了七萬里拉，漢城的朴教授幾乎被搶…。這是甚麼世界？回到了台灣，還在電子郵件上讀到加州來的回信：「…。你們能全身回國真好，我曾在那不勒斯被搶，損失 1000 美元…。」。

但這些都擋不住遊客如潮湧來。馬蹄清脆地敲在石版路上。這是三步一個古蹟、十丈一個驚喜的迷城。才出了我們住的名叫做 Brunelleschi 的旅館，就看到兩個日本團在導遊的指揮下猛按快門。過了七天才知道 Brunelleschi 也是一個了不起的文藝復興時代的藝術家，而這個旅館，有一部分是古時的女子監獄，是翡冷翠最早的建築！難怪它有一個相當古樸的石堡。

才從六米的古巷道轉出來，五十公尺之後，赫然聳立在面前的，又是教堂、鐘樓和另一座教堂。這是翡冷翠的觀光及購物中心。遊客一波一波地湧來，有的是大把大把的，有的是精打細算的，可是都在花銀子。

看過米蘭的大教堂，眼前的雖好，但並不那麼逼人。教堂用白、綠二色的大理石鑲出，有一些阿拉伯的味道。紅色的大圓頂是 Brunelleschi 的傑作。隔壁的八角型教堂建得較早。三扇銅門，最老的是 1318 年建的，一個世紀之後，一個叫做 Ghiberti 的人，在艱困的比賽中擊敗了 Brunelleschi，得到做另外一個銅門的權利。這個人一點都沒有偷懶，他用了二十七年。這個門後來被叫做 “到樂園之門”，因為米蓋蘭基羅說過，這樣的門，就是放在天堂的門口都配得過了。比賽輸了的 Brunelleschi 便全心放在建築上。他的精心傑作是紅色的大教堂穹頂，用八條白石撐起頂端的大理石燈架，內部是拜占庭風格的馬賽克裝飾。這是翡冷翠的象徵，任一個角度都可以清楚的看到。

這些是無數的歲月，無數的天才及汗水，再加上和宗教和藝術的熱忱所堆砌起的人間瑰寶：四個大字是「文藝復興」。江山才人哪，我們低迴讚嘆，渾然忘卻皮包可能被搶。這是美不勝收的所在，也是讓速食文化無地自容的所在。路邊盡是咖啡座。你要卡波奇諾，還是艾克斯柏索？一個黑人用心地為另一位在編極其麻煩的辮子，在往來的匆匆裡，添一些閒適。

無數的人在按下快門。留下的是雪泥，還是鴻爪？但這不是僅僅到此一遊的地方，它應該一寸一寸地讀。我們不久就學會了坐翡冷翠的公車。他們有很好的路線圖，略用心一點，雖不懂義大利文，也可以猜得差不多。

坐 13 路可以到 Pitti 宮。這是個美術館外加一個極大的花園，Pitti 少不得是當時的豪門貴族中的豪門貴族。不免的是朱顏已往，只有雕欄玉砌猶在。亞諾河將翡冷翠一分為二，自這裡過 Vecchio 橋，這個石橋沒有欄干，欄干是一家一家的精品店，多半賣義大利的手飾，像 k 金鏈、手鐲。手工精美，設計典雅大方，是天下最不能帶太太或女朋友來逛的地方之一。古來橋就是兵家必爭之地，今日商戰如古，橋仍然是必爭之地。義大利人可算真是在這裡佈下了重兵防守！

過了橋，一條街之後便是 Uffizi 美術館，對我們來說，它的有名是因為不久之前才被暴徒炸彈攻擊，上過電視。這是一個精品名畫幾乎不計其數的所在。我們有一個畫家朋友一連去了三天。

真正是一時消化不了那許多，文化上的不接頭，時空又那麼錯置。走著走著，迎面就橫著一座漂亮的古建築，多半還有人在用著。大街小巷裡幾乎隨時都有未先期安排的驚艷。那一天下午，我們走得累了，忽然聽到舒人心懷的風琴聲。尋聲問問，我們找到了一間教堂。這是個名字還不夠列在觀光地圖上的小教堂，但十分精緻。又乾淨、又溫柔。祭臺、十字架、雕刻，都一絲不苟。悠揚的風琴是今晚在此演奏者的預習。听和我找一個位子坐下，一句話都沒有說，但氣氛好極了。這就是在翡冷翠可能遇到，而在臺北怎麼也遇不到的。

自市外的米蓋蘭基羅廣場看出去，翡冷翠的古雅彷彿仍然是三五百年前另一個朝代的景色：紅屋頂、尖鐘樓、綠樹、長河、石橋。因此，當我們聽到鼓和號角，又同時見到一隊武裝的中古士兵沿著石板路迤邐而來的時候，我們也沒有十分奇怪。長戈大戟和盾牌鎧甲，原是來搭配這些尖頂古堡和滿街彫像的。隊伍走進市政府，市政府大樓又是一座古蹟，而葡萄酒及點心已安排妥貼。上千

的客人，正都熱切地等著市長致歡迎詞。

我們拿出受邀的請帖，走入古堡式的市政府，同時也和中古的武裝士兵用現代的傻瓜像機留影。這是十分翡冷翠的夜晚，廣場上的噴泉不息，古堡裡被邀的是從世界各地來的統計學教授及政府官員。真是十分奇妙的夜晚。踏著月色和石路，經過古巷回去的時候，仍然是滿街的行人。夜夜無夢，因為實在太累。

斜塔及其他

義大利有匹薩及比薩。前者是食物，有些中國人認為那是馬可波羅對餡餅的仿冒，他們只是把餡放在外面；後者是斜塔，中學物理就學過：伽利略用兩個球，一個一磅，一個五磅，同時自由落下。實驗證明的是，亞里士多德的理論錯了。後來的故事不必多說。現在的事實是，斜塔仍然未倒，仍然在每年多斜一點，而國中生都知道自由落體的公式。

斜塔邊上是白大理石教堂。比薩城昔日也風光過，1300 年之後地位漸低，但大教堂仍然一點也不馬虎。斜塔因為造的時候，地基不好，未能照預定的計劃造直，反而更為有名，這是啥道理？

大教堂裡有一個吊燈。導遊說，伽利略當年就坐在這個吊燈下，看著它擺來擺去，好像每一擺總是用同樣的時間。這個發現，使得他深思，並最後導致斜塔上的有名實驗。這些都遠在牛頓之前。那時節沒有國科會的獎助，沒有所謂的學術論文，異端邪說者反而可能被判火刑。是怎樣的上帝在那一個瞬間，決定了讓他領悟？人類由這裡窺看到古典力學和運動的影子，他們一點一滴的自這個「彷彿若有光」的小孔，抽出並精鍊成現代科學的基礎智識。走進這一步對科學的影響，真是不能形容的大！我靜靜看這個吊燈，今天它不動如老僧入定，誰知道它是開啟近代科學的最重要的鑰匙之一？

歸途在沿海的一個小城晚餐，時間是萬家燈火的九點。經過另一個小村是普西尼的故鄉，回到翡冷翠已近午夜。我們在翡冷翠的火車站搭錯了一路公車，而這已是幾乎最後一班。我們下車，車站上還有一個疲倦的埃及年青人在等車。這是來打工的外國人，他知道還有一班可以載我們回火車站。而自那裡我知道怎麼走回旅館。後來我們幸運地招手攔下一部空計程車，跳錶還是極快，但這次听和我都沒有抱怨。因為如果這是在瑞典，可能我們整夜都攔不到車。

水都

計程車東繞西繞，二十分鐘左右，到了翡冷翠的雷佛拉地站。這是個和南港站差不多大的火車站，精緻又乾淨，幾乎和五星級的旅館的門廳差不多。自這裡可搭自羅馬開來的快車，我們買的是二等票。火車座位和自強號相似，但可自前面座椅背下架起小桌檯，和飛機位子的道理一樣。火車既準時，又開得飛快，經過了好多山洞，好幾個大城市和一片大森林。這可不是甚麼高鐵，只是普通鐵路，一路上我們只有比汽車快的份。真不知道我們的國光號會比自強號要快是為甚麼？我們的高速公路可要擠得多！

威尼斯是最後一站，火車最後走進一條長堤，兩側都是水，你就知道到了。下車不久，旅館派的人舉著我們的名牌找到我們。車站外面沒有公車或計程車。前面只有碼頭，碼頭上來接我們的 Taxi 是一條二十多尺的漂亮快船。年青人在運河裡東一彎，西一拐，最後到了寬廣的水面。三十分鐘左右到了旅館。我們才知道我們住的不是在威尼斯，而是在麗都島上。

麗都島可算是威尼斯的貴區，可不是，汽船接人，是十二萬里拉，合台幣兩千圓。首先要學的是乘水上巴士，自麗都到聖馬可廣場，2500 里拉，旅遊資料方便而又非常翔實，用一點心就弄懂。弄懂也有不好處，我知道我們會錯過九月五日的大運河賽船，也知道趕不上威尼斯影展。

這是一個花錢的城。買東西不錯，熱鬧的街和臺北東區一樣擠，但好處是一輛汽車都沒有。船，各種各樣的船，是唯一的交通工具。所有的店幾乎都是漂亮的店，幾乎清一色的「義大利原裝」。這是有藝術天份的民族，不論是衣服、手飾、皮包、皮鞋，都有模有樣，大方得體，顏色工料，無不高明。這些店集中在幾條狹小的巷道裡，石子路、大大小小的運河、石橋、碼頭，都擠在滿是鴿子和遊客的聖馬可廣場四週。

古蹟也是很多，多半也記不得了。無非是歷史上的打殺，宗教上的膜拜。但知道這是個年年下沉的城，水位高的時候，冬天的時候，日子一點也不好過。建在運河邊上的房子都是幾百年的老房子了，不止是小，更談不上現代的設備。夏季的熱鬧是遊客和他們錢帶來的。冬天的寒濕來臨的時候，想必要寂寥得多。夜色裡，我們去坐獨有的 Gondola 船。靜靜地划，穿過大街，小巷、石橋。前船傳來的是散塔魯其亞那樣的歌聲。

夜涼如水，歌聲配著吉他、手風琴。水色和月色構成的是奇奇怪怪的異鄉之夜。運河邊的房子，二樓以上的燈光多半是不存在了，因為沒有人住。這不是一個活躍的城。白天的繁華，忽然變成了蒼茫的淒涼。

隔著水面的麗都島是另一個有錢有閒的世界。乾淨清爽但甚麼都貴。海灘是義大利最好的，但多半屬於私人。我們來了，又走。好像沒有深刻的感覺。我們在此用了不少錢，但離開的時候，就輕輕鬆鬆地離開了。這個城沒有引起我甚麼共鳴。雖然，我完全不能說，威尼斯不好玩。

回翡冷翠的火車是六個人一間小房的那一種。面對面對坐著，大眼對小眼，無聊得很。到站的時候，大家似乎都有一點如釋重負。再回 Brunelleschi 旅館，心情還頗愉快，雖然，那家麗都島上的旅館，早餐要豐富得多，咖啡尤煮得香。

那不勒斯印象

印象中的那不勒斯是「愛的教育」中的那個熱那亞少年眼中所看到的那不勒斯。藍天、海港、白帆。那一天，巴士在六個小時之後將我們送入鬧區的旅社，我們的印象，卻好像在紐約的一百二十幾街一樣。

南方的義大利是比較窮。窮而並不把衣服洗乾淨。雖然，衣服的料子和剪裁都仍是上品。這是真正的義大利麵，「司撥蓋第」的發明處。我們是受到良好保護的觀光團。有全程陪伴的全陪阿巴小姐，快樂而瘋瘋顛顛。到每站又都有當地的地陪，負責解說古蹟名勝。我們被送到當地最高的一家旅館，窗口望出去，景觀棒透了，可看到大半個那不勒斯灣和遠方的卡不里島。

如果減掉它所有的古蹟，那不勒斯就應該和三五十年後的台北一樣：過多的人口、過小的街、普遍地不守交通規則，因此人人都有高超的駕駛技術。旅遊書上的文字極為精彩：

…別在那不勒斯開車。太瘋狂了，簡直瘋了。第一天你應該用走的，或者坐巴士。但一旦你嚐出這城市的味道，就去品嚐一下那不勒斯交通的無政府狀況吧！單行道毫無意義，紅綠燈純為裝飾品，在這裡你學到的是生命中重要的一課：探測可能性的極限…

…在這裡有兩種對遊客們可致命的態度，旅遊指南繼續說明：一種是將那不勒斯人都看成唱「可愛的陽光」的可人兒；另一種是將他們都想成埋伏在轉角處想割你咽喉的賊。…在這裡，主要的藝術，是活下去的藝術，用最大的智慧克服低迷經濟所引起的困難，因此對於感性的情懷，就所餘不多了。一個本地的社會心理學家指出：帶有一點妄想症的遊客基本上是以他們過分神經質的行為向來扒手和搶皮包者呼喊：「向我下手！」而由此滿足他們的懷疑。有些人是天生的受害者…

旅館外就有大批的真槍實彈的警察或是士兵，我們只去隔壁去買礦泉水。夜已低垂，我們一點要去街上逛一下的意念都沒有，何況，八點半才開始的晚餐十點都結束不了。

海港應該是老歌「聖塔魯其亞」所描繪的。中學時學過的歌詞，似乎不那樣搭調：

黃昏遠海天邊，薄霧茫茫如煙，微星疏疏幾點，忽隱又忽現…

夜已深，欲何待？我們坐困在二十五樓的大廈，看碼頭、看燈火、看山、看海。這是散塔魯其亞沒錯。

龐培，凝結的古城

快門捕捉的是一瞬，龐培古城便是上帝的快門所捕捉的二千年前的一瞬。只是沖洗上帝的軟片不易，我們看到的，還是不全。即使這樣，已彌足珍貴。這裡，只要抬頭找一找，就看得維蘇威火山。葡萄是這裡的大宗，因為火山灰對葡萄特別好。

沿著 A3 公路，二十公里之後，我們來到這有名的古城。古城不止是頗有規模，且保護得十分完美。火山爆發是公元七十九年的八月二十四日。不久，羅馬人就幾乎忘了這個城，直到十六世紀，才因為做土地重整的工作而發現，其後經過小心的開挖，至今尚未看出全豹。街道、行人道、商店、市政府、市場、劇院、寺廟、浴室、旅館，甚至妓院都歷歷展示在眼前。

羅馬人是了不起的工程師，龐培必然是一個有都市計劃的都市。我們甚至看到原始的輸水鉛管，後來，羅馬人才放棄鉛管。他們付了很大的代價才知道鉛有毒：羅馬人的壽命很少超過三十五歲。不知道為的是甚麼，龐培人，依考據約二萬五千人，會受到這樣的處罰。若是說他們因淫亂而觸怒了上帝，那麼羅馬人也不見得更好。但是到了這裡，效果和進入時光隧道一樣：兩千年恍如一瞬。

晴空白雲，我們在千門萬戶中徘徊蕩漾。那時候磚和馬賽克已是標準建材。有錢大戶的門口，用馬賽克鑲嵌出「歡迎」的字樣。不知道兩千年後的人，是否還受到歡迎？離開龐培，也離開維蘇威火山。其實公元七十九年，被埋沒的還有一個叫做 Herculaneum 的城（它不是被火山灰而是被泥漿所埋）。但龐培太有名了，我們的行程裡排不出另一個時間。火山仍然是活的，上一次的較大噴發是 1944。自 1631 以來，至少有三千人因火山而喪生。這些旅遊資訊，不久就會忘卻。我們只是在時光隧道前快地瞧上一眼。

蘇蘭多海岸

這裡是美麗得讓你屏息靜氣的峭壁、懸崖、海岸和橄欖樹。山路蜿蜒，一下子，前面是迎面而來的山壁；一下子，轉角處赫然兩三椽紅屋頂，外面稀疏地在紅花綠葉間擺下幾張露天咖啡座的小旅館，然後幾千尺下面的海水蔚藍，白帆點點。山套著山，峭壁連著峭壁。葡萄園，橄欖樹。

白浪只為下方的海岸綴上一點細邊。旅遊巴士又非常湊趣，義大利女歌手自“散塔魯其亞”唱到“歸來罷蘇蘭多”。這是個極好天氣的日子。我們除了讚嘆，無話可說。

山路鬼斧神工，這是那不勒斯南方的蘇蘭多半島，義大利最羅曼蒂克的海岸。這是美人魚用曼妙的歌聲誘讓尤里西斯的船撞上岩石的所在。千年以還，情懷似乎依舊。峽及谷交錯著，這裡曾只有海盜及綠林豪客的足跡。但是野玫瑰和山茶點綴著岩石，檸檬、葡萄、栗子樹交錯有致。偶然見到一大叢怒放的九重葛，依稀是家中盆栽的舊識，便掩不住萬里之外偶然相見的驚喜。

沿著這條公路有好幾個小鎮。都是既靠山，又靠水，但既沒有沙灘，也沒有腹地。這些都是絕對的山城，幾乎找不到籃球場那樣大的平地。義大利人還是在做他們全年無休的旅遊生意。

我們匆匆來，匆匆去，知道世上有太多的良辰美景，而韶光易逝。

卡不里島

自那不勒斯的一號碼頭坐水翼船，四十分鐘便到了卡不里的唯一的碼頭。這是一個二十公里長的島，遊人如織。這個島也許除觀光之外別無大用，面積本來就小，又沒有平地，甚至也沒有像樣一點的沙灘。經濟雖有一些土產，例如音樂盒子，精美絕倫，但談不上任何規模。但是氣候好得要命，風景美得讓人心疼，是度假的地方，也是度蜜月的地方。

可以落腳的地方，都是美麗的花園及旅館和狹小的街道。這些都不記述了，反正都是昔日的達官顯要驕奢淫逸的所在，讓你覺得有錢有勢真好，如此而已。沿著峭壁建造的行人道和岩石、樹木、花草及石彫搭配在一起。近景是山風，樹濤；遠看是絕壁，海色。而遊客像螞蟻一樣地，沿著有限的幾條路線，跑來跑去。

絕美的風景應該只配上三兩個高士，眼前的景色是絕美的，但配上了上萬的高士，以及成千的小僮。有人大把花錢，有人大把收錢。構成了如此一個非常世俗的海上仙山。

值得一記的是有名的藍洞。用蚱蜢舟載人而入，遊客必需平躺在船艙，才能不受岩石括傷地，穿過這個彷彿若有光的小孔。裡面是有光，海水是夢一般的青青藍藍；你忽然發現你在一個五十多公尺長的岩洞裡，迴聲和船影，奇妙地給你意想不到的驚喜。但這樣的奇景，不可避免地變成了重要的賣點。在那裡討生活的漢子，長年的在夢幻美景與現實的海浪之中輾轉來去，也許聽膩了外來客的驚嘆。當他們手拉著岩石上的鐵鏈將小舟帶入的時候，心情也許還低於我每天推開辦公室的大門罷？

羅馬假期

如果咸陽尚在，阿房宮無恙，也許中國也有一個城可以比一比羅馬。羅馬是一個強大帝國幾乎在最輝煌的時代留下的記錄：莊嚴、雄偉、殘暴、勇武；是智慧和財富的結晶，也是權力和血腥的展示。英雄何在？英雄尚在，英雄無所不在！這是歷史的櫥窗，一個民族的偉大史詩的斷簡殘篇。經過了二十個世紀的洗鍊：用血、用火、用戰爭、用文化。這是讓人低迴不已的城，這是讓人處處震攝的城。讓你懷想，讓你詠嘆。羅馬就是歷史，就是氣魄，就是詩，就是夢。

我們看不了那麼多的羅馬：但是那是七山，是她的源起；那又是墨索里尼的總部，是半個世紀前的屈辱。你只知道通羅馬的是條條大路，你也知道羅馬不是一天造成的，於是你自這一個古建築看到另一個古建築，瞻仰的是英雄的遺跡，懷想的是奴隸的血淚。自由人歡呼的時候，也是奴隸、鬥士、教徒、獅子流血的時候；也是當權者分配利益的時候。

用長矛及盾牌組成的是有名的羅馬大軍方陣。戰車是兩輪的，由怒馬牽起。踏平的是反抗，帶回的是征服。這是天下的聚斂，這是用無數的沙場枯骨和春閨碎夢建成的蘭若。征服之後是彰顯偉大勳業及酬勞帝王將相的鉅大工程，這個模式是何其熟悉！

眼前見到的是離宮嗎？是迷樓嗎？是長城嗎？是運河嗎？還是圓明園？戰爭是為了權力，權力可以聚斂，聚斂可以酒池肉林，可以收納美女。那時候，華夏就是全世界；在另一半，羅馬也是全世界。以為擁有全世界的人，做法都差不多。我們走在歷史的路上，自公元前走到公元後，自沒有天主教走到都是天主教，自羅馬人殺教徒到教會燒異教徒。觸目的都是秦磚漢瓦、紀功碑、銅像、石雕、華表。有些是很古的，比耶穌還早；有些是很有原創性的，到如今還是

某一類建築物的唯一模仿對像；有些是碩大無朋的，建的時候就一定要天下第一。這就是看不完、走不盡的羅馬。沒有甚麼可以形容，因為他給你的，就不是一個可以想像的世界。我們只能在阿房宮賦裡，在長恨歌裡，想像一鱗半爪。

走兩步，見一個噴泉；再轉一彎，是另一個廣場。那是最早的超級市場，那是賽車的所在，那是皇宮，那是教堂。最後我們穿過古城牆，走到最漂亮的現代街道上。兩側是第一流的精品店和雅緻的咖啡屋，迎風招展綠雲的行道樹。那些風，那些月，想必仍然和古來一樣。

回到設備極為現代的旅館，我們走完一個二千年的旅程。腳踝已疲，昕和我。我們走過半個世界，半個歷史，我們依然是過客。羅馬假期已結束，除了清風明月，誰都是過客。

十二、山中日記(1968.2.2)

六月十日 晴極了

出了海華，車子向五十號公路的東部走。美國的好公路只有兩種，這一條當年也算是有名的大路，只是更新的超級公路不斷地建，這種當中用花木隔開，一邊兩線的快速道路已漸成二流的道路了。天氣極好，沒有雲的天上是一望無際的藍，車窗裡吹著的是舊金山灣這一帶的海風，油箱裡滿滿的，腳底下的油門感覺到的是那幾個汽缸裡鼓足的勁。

一路東行，五十號公路轉到一百二十號，路漸行漸小漸高，加州特有的西部荒原漸漸被紅木、青松所取代。山路裡沒有「鳥鳴山更幽」的味道，現代文明用尖利的刀斧在跡近洪荒的山地裡深深地烙上一條既長又平的帶子。熊、鹿、印地安人難得見了，便有也在那僅有的保留地、國家公園裡。而那一條文明的烙印又每天不斷地送許多受夠了文明薰陶的人去體驗那些遙遠的自然。自然是離我們遠了，從燧人氏第一次取到火種的時候，人類便不再臣服於自然。有時候我奇怪我們在做甚麼？但我們實在太忙，是嗎？瞧著隔座的她，我們已差不多開了五小時的車，路牌上所說的七哩崎嶇已差不多走完。快到了，我說，你能夠

替 Yosemite 找一個翻譯名字嗎？像翡冷翠那樣美的名字？

「幽山迷地」怎麼樣？她說。我反正是同意的，國家森林的看守人（ranger）收了我們的入園費，給了我們地圖，於是我們的車子向山谷裡而下。山谷清幽，山水清冽，但那似乎是第二天的事了。我們所得到的是五小時的疲累，滿天的藍，圍繞著我們的峭壁飛瀑，奇怪的石，青青的松。踩到的樹皮松果鋪成的森林地面，滿地的松鼠。有許多領略得到的和領略不到的另一種清醒。

終於入夢，我有一個記不得的夢。她也夢到了熊，也許是鹿，也許是我。

六月十一日 晴極了

幾年以來第一次被鳥聲泉聲吵醒，帶著城市人在週末的懶散，直到太陽落在帳篷上耀著眼睛才醒。夜裡奇冷，我有一雙下半夜冰涼的腳以及滿床壓得又厚又重的氈。

勤快的美國人已然大多去尋幽探勝去了。說也可憐，他們好像在哪裡都是同樣的一套。全美國的三文治、霜淇淋一個味道，是不是受了工業上成批大量生產的結果？一部車，開到東，開到西。整個美國好像都是為開車的人設計的。且不論那許多真正一流的公路，市場若沒有停車場，那麼便簡直門可羅雀。無數的小吃店、電影院叫做「開進來」（drive in），每一個稍好一點的幽點（原字為 view point），都畫好了可停車的格子。這不是騎著小驢踏雪的國度，大家太忙，驢子，你知道，沒法子一小時跑六十哩。

聽說最好的幽點叫冰川點，從營地裡仰著頭，直到帽子快掉的時候，便看得見那個點了。直接的距離大概只有一、二哩罷，對著營地的山都是峭壁——這也是為甚麼這裡還有許多瀑布之故——車子繞著上去，整整走了三十五哩山路。

山路有山路的美，幾個人合抱的紅木青松佔滿了多餘的空間，不是秋天，沒有楓葉，山間是一色的綠。偶而一、兩株枯了的松，背陽的一面長滿了綠油油的苔，鮮嫩的綠。多想見一兩頭啃青的山鹿呀，但所有的木牌都說這裡所有的鹿、熊都是野的，不可以餵，不可以逗，不可以騙它們。

據說所有的美國國家公園裡都沒有廣告牌，至少這裡是對的。山谷的高度約四千呎，四周的山約八千呎。顯明地感到車子沒有馬力，因為高空裡氧氣不足。殘雪零星，更多的白在更高的山裡，藏在更深的松林裡。一點點融，一點點化，匯成小溝，聚成各式各樣的湖，合成溪，聯成澗，涼涼地，潺潺地，流過松根，這些雪水染得整個山谷都是青蔥蔥的。輕輕的潺暖漸漸成為沉重的轟發，忽然之間，從那一點，前面是一落四千尺的懸崖斷壁。於是這些已變為綠的雪，又得變成倒掛的白。驚愕的水滴被山風化作薄霧小雨，四千呎的落下，把驚慌的，才見到人世的水變成憤怒的衝擊。但是峭壁有峭壁的頑強，怪石有怪石的個性，幾曾見過那樣的激盪，那樣的洶湧？這不像海，海太世故；這不像河，河太受到拘束；這是落在亂石叢中的飛瀑。它不夠強，因為沒有太多的水；但它夠傲。傲，懂嗎？這些是初生的水！

沒有人知道它們成為海還要多久，也沒有人關心這些。水自己也不會在乎的。我們在今天在所來的最好的兩個幽點之一「洗而燒」(Washburn)之上，俯視廣大的山脊。源遠流長的瀑布頭，山風勁急。沒有在阿拉斯加上空所見的雪山千里的雄偉，但眼前的景色是另一種強壯。藍色的天，蒼色的岩石，坎在石縫裡的松。頑強的石，更頑強的松，生命瀰漫在整個空間。石頭上奔躍的松鼠，山頭上的鷹，松間的翠禽。山風吹刮，我們在八千呎的山頭。

再過去是冰川點，已見不到昔日的冰川，據說是冰川的遺跡尚在。不懂得這些，也不理會這些，在這裡俯視的是整個幽山迷地的山谷。幾人合抱的松像小草，漂亮旅館的游泳池只是一為翡翠的碧，對開的公路像阡陌小道，穿流山谷的河水蜿蜒在青松岩石裡。

有許多東西提醒你這是美國。賣霜淇淋熱狗的亭子，給你同樣在舊金山能買的味道，旅館的櫃檯上你能找到那點型的堂信臉，買東西你得付百分之四的稅，仍然有山下印好的明信片，日本造的紀念品。

但那是另外一部分，也不是我們預計的那一部分。三十五哩的山路使我們下午四點才回到營地，而在九點半鐘的時候，當我們走出營帳，仰視自冰川點推下的「火爐」所造成的「火瀑」時，也忘記了那是美國商人的廣告花頭。而那時天色已晚，山谷裡充滿了營地裡升起的炊煙以及烤肉的香氣。租腳踏車的，租馬的，釣魚的，游泳的，划船的，紛紛回去。暮色來得很快，太陽一下子便躲在山岩的後面，滿山谷都是暮色。而在燈火紛紛舉起的時候，還有幾個山頭耀著閃亮的金黃。不知不覺，涼氣重重，而這是六月。

六月十二日 晴極了

在公園的西北方是一片高高的草原，導遊錄上說這是個該去的地方，去便去罷，誰管它有五十五哩山程？

五十五哩山程便是開車兩小時半的另一種說法。路行漸高，不止一次我們停下來找積雪的殘潔，熊的足跡，湖上的藍。在青蔥的松林裡，我們找到去年的雪。草地是濕濕的，雪在化，水在流，於是我們忽然好像找到了這些飛瀑的源頭。

不理會這些涓涓，我們辭了這些往年的白，無數的小蚊子。草原在望，這是六千到八千呎高的一片大平地。

石山散佈在四周，明淨的一條河——恕我不記得，也不曾去記名字——穿過多石的草原，穿得這草原都是一片綠。石、松，這兩樣好像是這裡的特產。無數的石散佈在斜斜的長坡上，無數的松也在那那。流水潺潺，穿過那 1907 年的木橋，水清冽極了。她踏著一株斷松做橋，戰戰兢兢地走到河心。光了的腳觸到河水的冰趕快縮回來。三兩個人在下游穿著長靴釣魚，他們也該覺得雪水所沁的涼氣罷。

不像桃源，這只是一片自然，而人來這裡不是為了逃避甚麼。浮生沒有半日間，在這裡旅行比在家裡更累。彷彿來這裡的目的便是把多下的一點錢浪費掉。美國人從不諱言他們的闊，他們各自奮鬥，因此他們的腰桿挺得直直地，心裡面坦坦地，太難看到一個羞羞怯怯的孩子，更難看到一個羞羞怯怯的大人。想甚麼，說甚麼，沒有誰在逃避甚麼。他們忙得沒有時間去感覺寂寞，而在他們能感覺寂寞的時候，社會也差不多開始冷落他們了。

但這不是社會的錯。我們在松下石上聽松聽水聽石，遠方的天飄一片淨白的絮，不盡的藍，廣大的青。可惜還沒有楓葉的紅，雪的潔。這些並不是我們的，我們也不想。瞧一次，夠了；又不夠。也許我們還有勁來趕秋天的楓和早落的雪——那是有顏色的秋天。夢，早就沒了。我也不要活在夢裡，詩裡。

租一匹馬怎樣？可惜沒有那樣的福氣。沒上馬背，便不知道世路是多麼崎嶇的。路上的石，路邊的松針，柏葉，一逕用他們的搔擾來告訴我們是外來的不速客。總之講好兩小時的馬程，不到四十分鐘時便叫停了。她，我想，也許再也沒有興趣騎馬了。

然後我們去享受我們浸在雪溶的河水裡的香瓜。風輕輕地吹，樹影緩緩地搖。河水奔騰，那是一個下午，陽光灑在翻白的河水上面。

草原靜靜的，只有河水的聲音。似乎只有我們在，至少我覺得只有我們在。天地悠悠閒閒，忙的只是我們。然後有一剎那你只覺得那天上的閒雲和草地上的雛菊是相干的，然後連那一點也不相干了。我們只有滿天的藍，滿河的碧，滿山的青。不去想那連天的歸路，不去想那隔著那樣多的山，那樣遠的水，那樣古老的夢。

六月十三日 仍然晴，並且累

昨天累壞了我們，而今天我們想該走了。早上結清了宿營的錢，又幸運地在幽山迷地裡找到菜場。難得這樣早便走，我們預備的是九十九號公路南下洛杉磯，但臨走之前我們要去南方的大樹（Big Tree）。

大樹並不是一棵樹，至少我們見到兩棵。最大的叫巨人，大約是三千八百年；另一棵叫隧道樹，因為恰好樹底成一個孔，汽車便可以從那裡開過去。

這一段路很不好，經常有老樹橫攔在公路上。自然，倒在公路上的那一段是鋸斷的，從斷面上可看見原先的樹是多麼的粗，使人想起滇緬公路上大蛇攔路的故事。

也有一個幽點，但遠比不上前天見到的「冰川點」和「洗而燒」。俯看只一片莽莽的青和三兩隻斜坡上的小松鼠。熊應該是經常在這裡出沒的，因為警告不要餵熊的牌子特別多。

看不到熊和鹿，也許這只是遊人的時候；看不到那住了三晚的山谷，眼下只有青莽莽的一片叢山。來路隱藏在這一片青裡，三千八百年的老翁巍巍地在和風霜蟲雨苦撐，這也許是這裡最老的青，而它的身傍，左一叢右一叢都是細細嫩嫩的幼松和柏苗。再三千八百年又怎樣？賣紀念品的地方，任一塊化石都是好幾十萬年！但活著的老畢竟使人尊敬，無所謂勝利和失敗，無所謂生命和死亡，只怕是那一叢小小的綠，那裡便是很難消磨的生命。跳躍的松鼠，松枝上的青苔，飛翔的翠禽，還有那些數不清的即將來臨的生機，明年的雪，明年的瀑布，和永無休止的陽光。

也許我們等不到這些，我們想趕這一秋的楓，這一季的早雪。衰勁的秋風裡，也許我們更愛那河水裡的落葉。披著白衣的青，也許更翠更挺。

離去了，我們。青山不留我們，我們也不多留。便像那山澗裡的水，偶然變成雪片，飄落在這一片山坡上，留緣隨住。撥轉車頭向南，我們又把自己投入另一個萬山叢裡，但這一次是更接近——至少在表面上——資本主義的青山。出了公園門三哩路，看到第一個商業廣告牌子，於是我們知道我們又不是遊山玩水的閒人。又要上班了，她說。我微笑著，也不知道笑容裡藏著的是快樂，是惆悵，還是淡淡的鄉愁。

1968 年 2 月，《皇冠》

後記一：

她有十天假，而唸書的我是無所謂假不假的。結婚之前便許願帶她去幽山迷地，而結婚以後便一直窮著。嚮往這裡久矣，去了一次之後，覺得還應在初秋時再去一次。她說她有旅行細胞，去玩了一次看起來也健康些了。

共在那兒待了三到四天，沒法子去每一個點，因為光是小路，這公園便有七百

哩，而面積差不多有半個台灣大。但想來再待下去也只是看些不同的綠，儘管每一片綠都有每一片綠的美，但那是屬於蜣蟲搬米的故事了。我喜歡我的感覺，也曾盡力表現我的感覺。據說眼睛是精巧的光學儀器，但我的腦卻似乎不是彩色的柯達。許許許多多的事情只是感覺，一些不必深究的感覺。便讓它模模糊糊地存在那裡罷。有人說那也是一種美，我倒是不知道，也不在乎。而這種不求甚解是她最討厭的。

1966 月 6 月 21 日，不可來城。

後記二：

我們趕上的，是那時最富足而自信的美國。記得她最初（1963）在 Oregon 向房東老太太要房間鑰匙時，老太太很生氣，意思是，我們美國治安這麼好，誰用鑰匙？1968 後到了 New Jersey，又到了 NYC，才覺得台灣至少治安還不錯。1982 我們回臺，又趕上每年 8% 的 GDP 增長時代，真後悔沒去買自己的房子，反而去住公家宿舍。2004 年她又去美國的洛杉磯，看到了當年她讀 UCLA 時的那麼好的高速公路，怎麼會變的那麼差？——因為我們已看多了大陸的好多新造的好公路，和洛杉磯的那些坑坑疤疤的還非常擠的高速，全不能比。時間不饒人，也不饒國。

2021，COVID-19 疫情中的臺北。

十三、同學重聚記事(2010.2.1)

五月二十八日，遊覽車一早從中研院的學術活動中心開出。目標：信義路的師大附中。車上有良磯、安達、建元、和兆復這幾家。他們老遠從國外趕來。帶路押車的是田昕，她可是一女中的。她才辦過一女中的忠、孝、仁、愛、信、義、和等班的 classes reunion，我們這點小場面（沒辦法，我們可是最後的春季班），是難不了她的。

(more…後續段落已遺失)

十四、破而後立(2010.3.4)

武俠小說裏有一招叫做“破而後立”，意思是說某主角（多半是男主角，從來沒看到女一號如此，小龍女除外。女俠或女賊受傷，主要的目的是給男主角機會，不是讓她們增加功力。）被打得差不多快死，然後突破瓶頸，武功大進。

根據微觀經濟學的理论，任何有競爭力的事物，一開始就有少數人眼明手快，得以輕鬆獲利。然後很多人開始搶食大餅，紛紛跟進。最後市場平衡，弄得大家都只有蠅頭微利。

教育也如此。大學生滿街走的結果，就是大學生不再值錢。等大家都知道學位有用時，學位就快沒有用了，尤其是那些靠花錢投入就能拿到的學位。

要徹底打倒升學主義，只有一個辦法：用完全隨機的方法抽籤。這樣做，可以100%地做到“破”。

我建議的是“鋸箭法”。

鋸箭法是說有人中了箭，請外科醫生治療，醫生把箭幹鋸了，然後說弄好了，剩下的箭頭請找內科醫生。

誰說破就一定立？即使在武俠小說裏，99%的人物，都是破而不立的。

我又不是教育部長，能替他解決一半問題，還不夠？

十五、不夠智慧時不如不要智慧(2010.3.3)

MS 的中文輸入軟體，有不少問題。下面是我練習了兩個星期後，遇到的困難。

1. 冷僻字太多。常用的字，不過 4000 左右。再加 2000 較冷的字，就應該差不多了。太多冷僻字在上面，大大地增加了選字的麻煩，同音選字，是漢語拼音的基本問題。再因為 MS 的中文輸入法不外加四聲，有時甚至要在 200 字裏選一個。不信的話，試試去找一找單一個“醫”字！
2. 同樣的一個字，常常不出現在同一個位置。這是設計成這樣的，因為這個輸入法以詞為主。用 yiyuan，很容易得到“醫院”，但是用 yi，反而很難得到“醫”。單用一個 chang，則“常”不是在第四就在第二。但是用 changchang，“常常”就出現在第一位。它是永遠記得我的習慣，還是只在這一次記得，我還不清楚。但是經常要在不同的地方找同樣一個字，怎麼可能快得起來？
3. 有些常用的繁體字很難找到。

這是號稱是“智慧型”的輸入法。在軟體寫得還不夠好之前，多半是自作聰明型的智慧，電腦用它的小聰明和愚蠢的堅持來告訴你它認為的對你最有用的方案，是很麻煩的。例如我用“ruanti”，可以得到“軟體”，是簡體字；用“fanti”，就得到“繁體”，這是什麼邏輯？

《孫子》：兵聞拙速，未睹巧之久也。

十六、銀髮再學中文輸入(2010.3.3)

我其實曾經會過中文 I/O，那是 1995 左右，會過的方法是“輕鬆輸入法”。是吾友鄭紀倫介紹的。他的老哥是寫“1995 閏八月”的，是職業作家。紀倫說他老哥每天 8000 字，就是靠這軟體。法子不難，每個字基本上只靠字首和字尾，鍵盤和倉頡法差不多。那時還不算太老，練習了一陣，雖到不了每日 8000 的水平，也勉強可以用了。

可是“輕鬆法”是一個小公司發行的。到了 XP 的時代，就一直不能提供新的版本。我的辦法是不換電腦，這樣拖到了退休，同時也拖到了電腦非換不可。電腦換了之後，我就又回到了以前只會英文打字時代。偶然要用中文的時候，就靠手寫辨識法。這就慢了，女兒說，十個指頭當然比一隻手要快。其實手寫法還有另外一個問題：用電腦多年，有些字看起來是認識，但已經不那樣會一筆一筆地寫了。

決定了，年紀大了，再去記住一套鍵盤碼，並能記憶到不用想就能信手打出的程度，並不現實。英文鍵盤我是還記得的。雖不能到手揮五弦，目送飛鴻的程度，但還能馬馬虎虎不用想就知道。另外，也不能再犯以前的錯誤，採用可能倒閉公司的 I/O 軟體。因此漢語拼音型的輸入法，就是第一類選擇。

有個說法是：好的科技產品，應該是不用學習的。例如用腦波直接指揮，不用手，也不用手指。但現在只是 2010，離 2050 還早。所以對於中文輸入，還是需要練習，很多練習。練習需要基礎，對於漢語拼音的輸入法，最需要的是你對發音的掌握。

我年紀有些大了，沒受過國語推行委員會所弄出來的那一套訓練。簡單的說，就是發音不夠標準啦。“因”和“英”聽起來難道不一樣？“民”和“明”當然應該也是相同的。多年來，把我的名字“民德”誤叫成“明德”的，至少 100 次，也好像沒有什麼問題。如果是李艷秋，問題就簡單了。但是有些字是從頭到尾就讀錯的，我就靠網上的 Yahoo 奇摩字典。它的拼音包括漢語拼音，蠻有用的。一般我們都是用字典查不認得的字，在練習輸入法時，查的都是認得的字。不對，查的是我們認為認得的字。不要懷疑，很多字是很多年來從頭到尾就讀錯的：標準國語是很難掌握的。

另一個問題是軟體本身。MS 的這個漢語拼音軟體，好像不是對繁體字很懂的人弄出來的。例如“國”字，就排在很前面，你若要繁體的，打入“meiguo”可以得到“美國”。如果只要一個單獨的繁體的國字，是有的，但是在很後面。同樣的問題，在“醫”、“後”、“夫”、。。。都存在。有些不是簡體繁體的問題，例如“夫”字，我可不知道簡體怎麼寫。這也不是冷僻字或常用字的問題。“轎”這個一般字典都沒有的冷僻字，都排在“醫”前面（“醫”排在 150 字之後，不受重視得令人髮指）。而且，簡體的“醫”，我也找不到。網路上把 Windows 叫做“瘟到死”，不是沒有道理的。

練習了兩個星期，我現在也勉強可以上手了。仍然很慢，但我也不用上網和別人聊天。後遺症不少，臺北市的街道名，用的是漢語拼音，捷運也用，出門時

可以練習。上午去銀行辦事，行員的名牌上有個“玨”字。我便問，兩個王字怎麼念？她說“絕”。我心裏就在 que 裏 jue 想，用“雞”還是“妻”？有一次在捷運上廁所，心裏還在想，噓噓應該是 xuxu 嗎？這樣的日子，也不是那樣的無趣的。

2010-3-3

十七、如夢的夢德里(1964.2.3)

霧像霧一樣，白茫茫地擋住了一切。從車窗裡望出去，我們正在有名的「十七哩」上。對開的公路兩側散佈的是殘缺遍地的白沙。寂寞的柏樹只留下淡淡的影子在霧裡。

海上沒有白浪，起伏的海面沒有澎湃，卻多了含蓄；一兩行靜靜的沙鷗，遠遠傳來海豹的吠聲和腥氣。霧幾乎是籠罩了所有，不像寂寞，不是激盪，只是一陣僅能感覺的壓力。

五點半了，暮氣和霧氣更濃濃地混在一起，我們下了車，腳踏在風蝕的石岸上。海風很弱，濕氣卻重，一位中年婦人推著一輛嬰兒車走過。這裡四週有一種淒涼的美，她推著小車，從霧裡來，又從霧裡去。

一切有若夢境，這許多暮色，這一個下午，這遙遠的異國，這幾個十幾年的朋友。有許多落寞，有許多迷濛。真實離我們很遠，三五條稀疏的人影，更加深了些淒涼。異國，遊子，靜靜的海，漫漫的鷗。沒有目的地，卻只戀著這一片迷茫，等著走入下一個未知。

1964 作。人在加州苦讀。這短文刊載在某一期的皇冠雜誌，已找不出時間

十八、靜靜的明湖(1967.7.7)

車子爬完最後一段曲折的山路，兩小時的紅木和蒼松困著我們。眼前是無數的一二百呎的直林，困住我和她的是那一片無盡的綠和頭頭上一溜的蔚藍。讓奧勒崗州保持她的線罷。嘿。她讚嘆眼前的忽然一片開朗：美極了。

我們在海拔八千呎上。山頂像高高的鍋沿，明湖如一面極藍的大鏡，靜靜地在群山環抱裡呼吸。靜靜的，她在我們二千呎之下。天是一色的藍，但藍得不如湖裡的湖水；雲是幾筆渲染的粉，但白得不如山間的殘雪。一片明湖，靜靜的火山口，透人心坎的藍，沁人的野松香味，跳躍著的松鼠和小鹿。我們到了，她喃喃低語：又多美！

數不清的年代以前，大地搖動，怒吼。吼聲裡帶著火，火光裡飛出石塊、樹木，流出了岩漿，巨大的恐龍被掀起在空中。也許那時還沒有恐龍，還沒有樹，還沒有記憶，還沒有這樣的藍。終於大地吹破了一個口，二十一平方哩的口，那裡，洩去了亙古以來的牢騷和不平，飛去了那一萬呎的山頭。然後山是那樣的靜，自然又讓時間去替自己修補、增添。冬去，春來，花開，花謝。無數年代的積雪，一年一年化作了來年的春水；松代替了羊齒；松鼠、鹿、熊又代替了劍齒虎、翼手龍。靜靜的明湖，四圍高聳的山牆，她收集長遠的雪，天公的淚，替月姐做妝鏡，替這廣大的土地睜一隻澄澈的明眸。靜靜的，連微風都沒有，一面藍玻璃鏡子。

山風吹起她的頭髮，頭髮亂了。我愛那頭髮帶子的藍，也愛這滿天的、滿湖的藍。她的眼裡有我，這滿湖的藍裡有四圍群山的映影。我們在山上，因此也在湖裡。這便像這四圍不盡的青松、紅木、松鼠、小鹿。山風吹著，刮著野松的香，山野的靜，雲影的寧。

我們漫步在山牆的沿上。火山口的遺跡像大大的盆，盆邊是二千呎的大滑梯。自一株松樹繞到下一株紅木，逗這邊的一隻松鼠，又瞧另一株斷枝上的紅胸鳥。湖裡的霧島，眼前的石級。

青山、紅木。我們自那裡過來，晚間仍將伴他們一宿。天色漸漸暗了，涼意重了。藍色的眸子緩緩地合上眼簾。倒影漸漸淡了，巫島漸漸暗了，陽光留著留戀，替臨去的一日，更為它所照射的一片明媚，幻上一層暮色。小鹿在暮色裡出來，而我們在暮色裡向夜景告別。也許這是千千萬萬個暮色裡的一個，而對她、對我，卻是這一生最值得回憶的一個暮色。我們很滿足，更很快樂。星光漸漸上來，還有上弦的一彎。然後青松的影子，紅木的影子，山巒的影子又出現了。它們不再是青的松、紅的樹、而且它們還混雜著我們的影子。

1967 年 7 月《皇冠》

十九、南港微雨(1993.6.2)

解聘了兩位博士之後

討論已經兩小時，最後的結果已可預見。研究論文不夠多，評審人如是說；委員會如是說；背景資料也如是說。這樣的大事要慎重，程序要明白……討論反覆著類似的掙扎，而當事人的呼喊又如此地蒼白和遙遠。科學的精神，變換成論文的數目之後，卻也和職業投手的自責分一樣傷情。但是選票終久是要發下來的，並且終究是要做一些選擇的。於是，終究有了最後的宣布：有史以來第一次因為研究論文不夠的未能續聘。

那意味著甚麼呢？是三十年為儒冠所誤的幻滅，還是終於脫去袈裟，又拾起屠刀的覺醒？經卷不外乎法華、秋水；禪意不外乎大屎橛，小屎橛。同樣的暮鼓晨鐘，怎麼會有的拈花，有的無淚？明日酒醒，又該是哪一個故鄉的風月？

從五個半小時的嚴肅會場出來，前面的難得清新的微雨。微雨如綿，滿地是南港的無情濕；我默然，心情如迎面而來的微雨。

1993 年 6 月 2 日，自立晚報。

二十、回首來時路(2000.10.2)

我在讀大一的時候，「大一英文」裡，就選了一篇幾乎每一個老師都會教的文章：「If I were a freshman again」。這篇文章其實我也記不得了，反正裡面的忠告，大概我都沒有好好聽就是了。今天我又有一個機會給大一新生說話，心裡想到的是這一篇文章，同時也在想，大概你們也會和我當年一樣地將它當作耳邊風來處理。人類反正總是一再犯同樣的錯。

如果我再回到大一，我是不是還會唸數學系呢？

那個時候，理學院還是有志青年的第一選擇，只有較沒有理想的、想賺錢的才去考醫學院和工學院，至於商學院，我們是不會去看的，想想看，他們只讀兩種微積分哪！甚至於商學院的女生我們都不那麼想去追。至於讀商學院的男生，當然清一色都是數學白痴，還有甚麼好來往的。孔子不是說：無友不如己者……？文學院的女生雖然數學更差，但起碼還知道雪萊、濟慈和一點存在主義，如果長髮飄逸，是馬馬虎虎可以商量的。數學系的學生雖然在系上被老師當得一塌胡塗，但是走在路上還是很有氣勢的。

這是一九六零時代的事。回到公元兩千年，好像甚麼都變了。

你們是甚麼分數進這個系的？

這話聽起來就不順耳，但是人應該聽實話：你們是因為高中的數學好才來唸這個系，還是因為高中的數學沒有唸好才來唸這個系？

數學這一行的確的風光不再。我全沒有在這裡騙你們永遠留在這一行不出來的意思，雖然，我也不勸你去轉甚麼企管，財務，會計或者c等等的學科。至少，不是現在就去轉。

一個擺明的事是：數學學得愈好，愈只能去教數學。最多是去教較好，較高級的學校而已。如果你真的很棒，在一個好學校做一個好數學教授，日子還是可以很快樂的。但是如果你真的那樣棒，數學能唸到那樣的水平，那麼你差不多做任何別的事，成功的機會和速度要比做一個了不起的好數學家要大得多。我隨便舉一個例子，在十幾億的華人裡面，只有兩個人做到了美國科學院院士，並且，他們一個是陳省身，一個是丘成桐，都不是臺灣培養出來的。他們的年薪當然不錯，但也並沒有十分地高。華人中名和利比他們好的比比皆是。

華人裡沒有其他的大數學家嗎？用世俗的眼光來看，是的。因為其他的人，就是沒有做到當選為美國科學院的院士或者沒有得到甚麼國際大獎。理由可以有很多，我們在其它的地方也聽得到，我們自己也常用類似的理由替自己找藉口，但是，大部分的人的解釋，就只有這樣簡單：沒有那麼好。在臺灣唸數學真的沒有前途嗎？那當然也不見得。我就認得一位在臺灣私立大學的畢業

生，最後一路奮鬥，能在哈佛大學做大教授的朋友。何況你們還是國立大學的學生。

做好數學難不難？當然難。其實做一個誠實的二流數學家都蠻難的。做一個三流的數學家就相對地容易了，只是他們不怎麼被二流數學家看得起而已。做一個不怎麼誠實的數學家恐怕更難。為甚麼？因為數學是一門老老實實的行業，俗語說：光棍眼裡不揉沙子。數學裡面，硬是沒有方法灌水。因為灌水絕對瞞不過同行的專家，而數學家的習慣，一向是不肯說別的數學家比自己好。在數學界要出名，只有靠硬功夫一途。口才、儀表、反應的敏捷、對人處事的態度，甚至幽默感都有一點用，但是這不是好數學家以及數學家成名的充分條件，也不是必要條件。搞得太多，有時反而有害。論文多也沒有用，好的數學家都不吃這一套。

我沒有向大家潑冷水的意思，但是要做數學家，甚至於應用數學家、統計學家、計量經濟學家、或者電腦計算這些以數學為基本工具的學門，想要佔一席之地（我不是說做甚麼長等等），硬功夫是唯一的路。我常聽到同事們抱怨現在的助理連程式都寫不出來，更別說做一點真正的解析計算。用這樣的態度在廣義的數學界想出人頭地是沒有希望的。

風光不再的艱困學科，由聯考の後段同學來讀，如果不能有一點想法，是蠻危險的。現在要從數學系（其實任何一個系都一樣）畢業並不難，如果你只是想畢業文憑的話，是可以放輕鬆一點，只要你不是十分的不上道，四年一晃就過了，我們的教育部，現在給大學文憑的態度差不多和「憑身份證領取」一樣，有甚麼可擔憂的？但是，想一想罷，你和你的家人，為了護送你進國立大學，這一路上投資了多少？現在可是滿街都是大學生的時代，數學系的畢業生，如果補習班都不肯要你去教，數學肯定不是第一流的，你又不曾管理，更不懂財務；會計？不會；程式，只會 EXCEL，別人會笑一笑：謝謝，再聯絡；英文？比不上文科的同學；國文，算了罷，大量用手機的結果，恐怕連情書都寫不好；聲、光、電、磁、氫氧、矽、塑一概最多只知道皮毛，不會超過牛頓雜誌的內容。誰要給你工作？

王爾德說：男女因誤會而結合，因了解而分開。國家設立了這麼多數學相關科系是一種誤會，他們以為科學可以救國，而數學若不是科學之母，至少也是科學的一種。最少最少，數學是一門不怎麼花錢的科學。我們設這麼多數學相關科系也許是一個錯誤，因為一個國家並不需要那麼多數學家。但這並不一定是無救的錯誤，因為，即使我們不需要那麼多數學家（何況訓練一個數學家又談何容易），但是數學的基本訓練，卻是極有用的訓練。一個人若是數學還學得可以，至少他或她的智商不會太差，邏輯清楚，能夠理性地分析出一個問題中的要點，知道甚麼是真的證明，甚麼是假的證明。而這是一個像樣的知識份子的重要條件。

那麼你們將要學到的數學有沒有用？極高的數學有沒有用我不清楚，大概「數學愈高深便愈無用」一般而言是不會太錯的。至少，從世俗的眼光來看，不會太錯。但是你在現在的數學系裡反正也不會學到極高深的數學，而那些最淺的、已充分發展的數學，像是微積分、線性代數、微分方程、複變函數這些

不止是真的有用，而且還是極為有用。在這裡，我是很世俗地說，所謂的「有用」，並不是課本裡常見到的「舉例應用」的那一種，而是有頭路的意思。

略為高明一點的技術性工作無不需要這些本事，你們如果在將來的工作中善用這些本事，已經和別人大大地存在著「立足點的不平等」了。可是，我在技術性的行業裡三十多年，見到的能夠善用微積分的人都寥寥可數，因為絕大部分的人，其實都沒有善用微積分所給他的技術。這和電腦的發達與否無關。不論電腦如何發達，仍然需要解析的能力。但是絕大部分的人，都到不了「能夠善用微積分」的水平。

不要以為這是因為老師教得不好。你們雖然不幸地到了這個系，但就數學論數學，你們在這裡遇到的數學老師可還真是上等的。你們若是學得不好，多半是你們自己的錯。但是反過來看，真的學好這些功課也並不那麼容易。它們雖然都是已發展成熟的學問，但是它仍然有它的基本難度。這些學問靠偷巧、靠臨時背一背是不行的。你必需要下真正的功夫。到了將來在職場上將這些學問活學又活用更難，因為你們到目前為止的讀書，都是為考試而讀，不是為了有一天真的要用。這些都是工具學科，你要能體會出這些工具是如何設計出來的、可以做那一類的事，換句話說，讀這樣的書要有整體觀。但是過去所有你會在乎的，只是那一個部分是必考的重點，那是讀補習班的態度，不是讀大學的態度。

我們才說到工具。真正的高手是不用工具的，他們用人，用懂得用工具的人。次一等的高手懂得看數字。但世上少有真正的高手，連次等的都不多。而且大部分的人在投入職場時，也不可能一開始就有用人的權利，也不見得會立刻看得到有用的數字。你總得先打幾個硬仗，才能從小兵一步一步升到將軍。而庸手是有工具也不會用、有數字也不會看的，大部分沒有頭路的人皆是此類。一個人在社會上有一點所謂的成就，其實只要能把握住一點利基，並且能發揮這樣的利基便好。你在一個數學系能學到的本事，是永遠都不會退流行的。

那麼，你在數學系的四年日子，應該怎麼過？

如果你是天才，不必問我，我也不敢給天才做建議。

如果你不算天才，如果你能把我前面所說的四門課裡的兩門，唸到至少可以去補習班教它的程度，那麼，你在數學工具上，就已遙遙地領先很多理工科的大學畢業生。並且這個學問別人搶不走，你終身有用。

其它的時間，試著去吸收其它的學問。學問可分為兩種：一種靠懂，一種靠背。靠背的東西要在年青時學，例如外國語。有的學問是要先背熟了，一背再背，經過一段時間之後才自然會懂。

例子：極限的定義。

需要時時更新的東西，除非你馬上要用，不要急著學。我曾懂過七種不同

的電腦作業系統，但是客觀的環境一改，你馬上就得換系統，以前的知識基本上就等於作廢。這類的學問是最好用時才去學。同理，少在網路上做言不及義的聊天，沒有用的。容易的東西、讓你爽的事物、大概都是沒有用的。

不妨去玩。但玩也要設法玩到差不多是職業的水平和品味。半調子除了做龍套之外，別無用處。

畢業了要做甚麼？

高手不必要唸更高的學位，王永慶、張小燕都沒有大學文憑，阿扁也沒有碩士。我們是普通人，因此人要金裝，需要文憑來證明我們其實還不錯。

下面的一段是選自「玩性堅強」一書的十八頁，作者是網路名人，筆名「老貓」，在文章裡自稱「貓」：

親愛的朋友，如果你想致富，貓在三十三歲前的建議是：

找一個有前瞻性的行業，別找夕陽工業，選個有錢又有能力，或者很會找錢但沒有能力的老闆，讓自己擺個忠誠又有能力的樣子，給自己裝個不貪財又值得信賴的德行，讓身邊的每個人都說你好，少交掏心挖肺的朋友，想盡方法認識其它更好的老闆，經過幾次半推半就的跳槽就可以了。

如果你要唸碩士或者更高的學位，那麼前面我建議你累積的功夫，可以讓你更有更大的可能，考進其它的研究所。用四年的時間，睜開眼睛好好看看你下一步要做甚麼。趕熱門也好，不趕熱門也好，你要確認的是：這是你的戰略，你深思之後的選擇。

從前，大學很少，大學生也很少。畢業生總找得到事，因此，專業不那麼重要。大學的教育，雖有一些學科學門上的差別，但基本上是以人格的養成為主要的目標，希望能夠培養的是有社會良心的知識份子。後來分工漸細，理工農醫文法各司其職，大學漸漸有職業導向的傾向。現在，這麼小的臺灣，有一百多所大學。大學生的平均市場價值當然也跟著下降，社會和他們自己對大學生的期望，也當然跟著下降。因此多半的顧主，大概也不會期望你們有甚麼真正的專業。也因為這樣的結果，也許大家反而可以忘記一部分對專業的期許，投注應該有的資源在人格的養成上。等到了我這樣的年紀，就知道專業只是敲門磚，成功的人都另有成功的特質。

回首來時，我其實並不曾給大家應有的示範。世事多艱，大學生活，是最後的快樂日子，也是機會。這對於自己覺得不幸考進這個系的同學來說，其實是一生中最好的契機。

如果你完全不愛解析、計量、科學計算……，那麼，趕快轉系罷。人生苦短，不要為任何事大大地折腰，這是多元的社會，路不止一條。不管怎麼說，興趣還是重要的成功因素。如果你還算能接受，並且還是覺得科技、管理、財

務、工程這一類工作是可以安身立命的，那麼，在這個系好好地留下來替自己的下一步做好準備，也許是不錯的戰略。

用心地玩，用心預備。只要你將來想要的工作裡需要計算、數量、以及邏輯——這是現代裡社會裡最要緊的三個部分——那麼，你應該不會吃虧，不會後悔。

我今天是來給大家打氣的，也是來給大家漏氣的。因此大部分都是講的老實話。老實的話不好聽，但是你往往在最關鍵的時候，才會被人提起。

電視廣告上，女孩的父母對來求婚的男子說：「你連房子都沒有，憑甚麼要我們把女兒嫁給你？」你早晚要面臨這樣的真實，尤其在面臨人生重大的轉折時，你需要拿出實證和業績。

在我們的時代，我們也沒有房子。但我們可以對女孩的爸爸說：我有本事。

本事不是一天就會有的，並且誰也不知道那一天會用到甚麼本事。

本事是信心，學識和勇敢面對。本事是效率：努力地玩，但要玩有深度和難度的事物。努力地積蓄真實的功夫，認識人性和自己。在自己的地位不高的時候不卑，在自己的地位略高的時候不亢。

我完全沒有許給大家一個未來美景的意思。日子只有愈來愈難過，競爭只有愈來愈劇烈，學位只有愈來愈不值錢，求得學位的代價只有愈來愈貴，貧富的差距只有愈來愈大。

但是，金字塔的頂端，總是缺人。我們找不到好的合作夥伴，找不到好的總經理，找不到好的部長，好的院長，也找不到好的總統。不止是臺灣，全世界都一樣。

有時間價值的東西，別急著去學。會讓你快樂的事物可以做，但要少搞。只是讓你輕鬆的事少做為妙。基礎的功夫，一輩子都是你的財富，要現在就努力經營。

數學系是少數可還以給你一點「永遠有用，永遠不會被搶走」的基本功夫的地方。它絕對並不容易，但至少基本的部分，也並不高不可攀。

最後，再引一段老貓的話：

如果你也是一個不屑的痞子，或立志當一個不同的凡夫俗子，你可以用貓三十三歲以後的法則：

有個專業，有個了不起的專業，有個別人沒有的專業。

搞好你的英文或者是日文，那表示你可以閱讀最新的資料去唬人。

找個明天的行業，別找今天的，這世界變化得太快，你總是追趕不及。

少說一點謊話，少做一點假，甚麼圈子都很小，你的信譽要用一生打造。

別相信專家說的話，相信自己，如果你懂得去聆聽心裡聲音的話。

少交朋友，小心朋友，因為那分好與壞，男與女四種，都很麻煩。

少睡一點覺，多讀一點書，沒機會看世界，從書裡也能體會，網路更直接。

多用腦去想，用眼去看，用心去體會。

如果這樣做，你可以吃得飽，睡得好，可能發不了財，可是你也不會窮途潦倒。照鏡子的時候，偶而還會感到自己順眼一點，運氣來了，搞不好會發得像豬頭一樣。

註：所有的引文，都來自老貓：「狂妄與不凡」一文，載在圓神的「玩性堅強」一書。

對高雄大學的應用數學系第一屆同學的講稿
2000/10/2

二十一、今夜星辰依然閃爍(2001.11.27)

寫一個我認得的國民黨大官。

汪鋹，字海儒。從科員做到主計長。

有時候還是不免有大官的樣子，但基本上是蠻負責任的大官。

第一次到馬德里的時候，覺得氣候乾得很。在機場搭上巴士，最後到了一個三顆星的旅館。大家都住在那裏：海儒公，周和刁兩位院士，黃局長，正之，麗雪，還有我年輕時就買過他的「統計學」的陳超塵教授。

那是 1984 年的 ISI 大會。海儒公是我們的團長。西班牙是非常不錯的可以去玩的國家，南歐的風情是紅色的土壤、蔚藍的天、懶懶散散的日子、同樣懶懶散散但熱情好客的西班牙人。

大會開得不錯，因為團長很罩。去開會的目的，因為是「排匪納我」，當然沒有辦成。但那雖是重點，卻又不是重點。真正的重點是後來便隱然地有了一個關於 ISI 的團隊，而這個團隊一連十五、六年都一直在為我國的國際統計地位在奮鬥著。

這樣奮鬥著，不知不覺地便走了好多地方。和海公同行，自然有和他一起走的好處。我國的主計系統遍佈全球，這時候便看出這個系統的深遠影響：我們一路辦事，同時也一路得到重要的當地支援。說白了，各條各路走來，到那裏都遇到的是他的老部下。那樣的熱情，看得出來的不是因為上級長官到了，而是一位可親可敬的老友被盼到了。於是那樣的異國夜晚，那樣的東南西北的風情，那樣的又中國又異鄉的沈醉，那樣的耳熱、酒酣，我們曾那樣地走過那多的路。

三十功名，八千里路。是誰說過？重要的不是結果，只是過程。

誰都知道他是一個才子：又詩、又酒、又橋、又畫；誰都知道他是一個第一等的公務員：有能、有力、有為、有守。讀過他的七十自述，便知道這樣的人，是因為生來就是這樣的人：只要有一點點機會，就會從重重的困境裏做出貢獻，脫穎而出：從那樣荒涼的河西走廊，從那樣艱辛的抗戰日子，從做得那樣一絲不苟的小代數習題，從這近幾十年不知開了多少次的無趣的會。

可是他飲來仍然十分阿沙力，唱起來仍然十分有韻味，跳舞的時候雖然大大方方地摟緊美女，拍子和步子卻是不會亂的。

認識海公是我非常重要的經驗，重要而值得懷念。從這裏我得到的是對主計制度和主計人的尊敬：前者並不因為我和主計單位熟識而增加敝所的預算，但也不虧待我們；後者表現在他們專業堅持和酒量，在我所參加過的這些國民所得的評審會上，他們極少不能明白說出某一筆資料的來歷。凡是主計處的邀宴，我鮮少能知道第四道菜是甚麼，因為「三巡」是雖然很快樂但是卻十分吃重的工作。作為帶頭大哥，海公在這方面所領導的品質和掀起的氣勢，是可以成為經典的。

欣逢海儒公八十大慶，我也借這個機會來記起往日情懷。我們曾很認真地在一起做過工，也曾很快樂地酒到杯乾、醺然忘我。問我何由醉？那只是好的感覺，尤其在回憶的時候。去年，海公得到了統計學社的「終身成就獎」，然後我們在德松的安排下去有如大峽谷一般的二零三高地。情懷如舊，而夜色漸漸上來。那樣的夜色是溫柔的，繁星滿天，四圍都是蟲鳴。

是的，繁星滿天。我們不知自己是那一粒，但是那樣的銀河，本身就是一種感動。我們也不在意，我們也在意。大家都在閃爍著罷，只是，帶著關懷的、帶著意志的、帶著人文素養的、帶著微醺的、帶著笑容的，就是有點不同。

2001/11/27

附註：題目用的是海儒公最愛唱的歌之一的尾句：林淑蓉的《昨夜星辰》。

二十二、Goddamn pandas(2007.2.4 轉貼)

We went to the San Diego zoo last weekend. There was a line of people waiting to see the Great Pandas. Once we got into the panda enclosure, there were zoo volunteers who were shushing the excited tourists. “Shh! Don’t disturb the pandas!”. We were also not allowed to linger as there was a line of people waiting about 100 ft. long.

We also went to the reptile exhibit. Here, there was a poisonous Bermudan snake that is critically endangered; it’s natural habitat has been reduced to 10 sq. mi. due to deforestation and development of tropical resorts. The only people looking at the snake were a bunch of punk kids banging on the glass of the poor snake’s aquarium. There were no zoo personnel to been nearby.

OK, so pandas are endangered. But (speaking as a biologist) these things have pretty much evolved themselves into a dead end. They don’t have much to blame for their endangered status except for their own evolutionary stupidity. The Bahaman snakes, however, can attribute the eminent demise to unabashed capitalistic human greed. I suppose I understand that pandas are much fuzzier and cuter than snakes, but I can’t help but think: goddamn pandas.

From <http://rustydomino.blogspot.com/>
This is my son’s blog

卷四 學術隨筆

一、我所知道的一點 Data Mining (2001.2.2)

前言：

Data Mining，光是看它的 data size，絕對是一頭大象。因此各家各派的對它的看法也就不同。因為各人摸到的是自己可以摸到的那一部分。兄弟雖然知道同時也極力呼籲 DM 應以實際有用為主，但基本上還是學院派。通常，學院派的人認為，如果一件事一定可以做成，那麼就不免無趣。幾十年前，我有一位朋友試著用海水提鍊鈾。因為那是用 ppm 來計算的，當然十分難。DM 有點像這類的工作，但是你不必只想造原子彈。提鍊錳、鉀、甚至淡水雖然無趣，但都頗有意義。這裡面的要點是效率。統計上所謂的大筆數據，和 DM 上所能見到的是不一樣的遊戲，因此雖然統計學者有明確的基本概念，但恐怕還是要因應變局，重新思考。反過來說，做 IT 的同仁，也不要一味相信背後邏輯不明的埋頭苦算會發現甚麼新大陸。總之，這算是新行業，大家各憑本事和運氣來博它一鋪罷！

今年十二月八號，我們在蠻辛苦的籌備之後，總算把 CDMS 成立了。這是一個以 Data mining (DM) 為專業的人民團體。在台灣恐怕是唯一的一個。DM 是一個蠻當紅的專題。我們所面臨的第一個問題，便是不知道應該怎麼去翻譯這個名詞。硬譯為「資料挖掘」雖沒有大錯，但聽起來不夠高雅，也沒有學術的味道，連商業的氣味都不夠。一時之間因為實在想不出一個好詞，只好馬馬虎虎地用著，等誰有了好譯名再說。我們想的事是把這一群人集合一下，看大家有甚麼想法，或者，更要緊地，看大家想幹些甚麼。

甚麼叫 DM？關於定義，各人的說法不一。基本上，先給你一個大的 data set —— 假設你可以讀它，下面的問題是：從這裡你可以得到甚麼？從這個角度來看，這是百分百的統計分析工作，只是 data sets 大得多罷了。但我們並不想忘記這些 data sets 是做 IT 的人這麼多年裡逐漸製造出來的。何況他們對於如何有效而快速地存儲、摘取和管理數據，確是有一套真功夫的。在 IT 的骨架上把統計的精髓放進去，的確是這個資訊時代的大挑戰。

極大部分 —— 如果不是全部 —— 的 data sets，是已經存在的。時代不一樣了，現在數據來得既多又便宜，多到沒有人有時間去看的程度。這有一點像我們的醫療體系。早先的醫生要望聞問切，在少量的數據裡，設法組合出最好的治療方案。現在則是一個名醫，一上午要門診到一百五十個病人以上。他只能靠 algorithm 來看病：問一兩個問題，聽一兩個問題，然後開藥。這是 DM 的精神，並不是傳統的 data analysis 的精神。

我們的確已經相當程度的資訊化了。你去任何一個醫院，哪一個窗口都是一架 PC。醫生暗示你可以離開的時候，當值的護士已在她的 PC 上鍵入許多資料。於是你直接可以去批價拿藥，並且連下一次的預約都安排好了。前天我去醫院調借一張 X 光片，一位五十多的老職員用一陽指在鍵盤上一步一腳印地敲了不知多久，才輸入了該有的 input，然後大家都鬆一口似地讓電腦去

跑，最後當然是拿到 X 光片。回顧身後，已排了一條長龍。

這在某種程度上，也暗示著目前資訊社會的問題：很多別的事情都還配不上那樣快的電腦。老職員的手指速度固然配不上，我們統計人和統計方法也配不上。甚至於很多統計觀念恐怕都配不上。於是 IT 的人就跳進來做：既然有了花大錢來做出來的 Data base，而且這麼多年下來資料堆得滿坑滿谷，好歹要告訴出錢的老闆「這有甚麼用」。DM 便應運而生了。

DM 是先有了數據才興起的行業。我不想說「學問」，因為到現在為止，我好像都看不到大學問。DM 能幫我做甚麼？不同的人有不同的看法：比較樂觀的是 Berry and Linoff(1997) 的說法：

- 分析報告給你後見之明 (hindsight)
- 統計分析給你先機 (foresight)
- DM 給你識見 (insight)

這話說得太強，我不是那麼地相信。因為這三者都是在既有的資料上做分析，在概念上應該並無軒輊，差別只是手上的 data set 大小和性質，因此，由方法的不同才有定義的不同。

較負面的定義來自 Friedman (1997)，他說 “Data mining is a commercial enterprise that seeks to mine the miners ”。這句話有多少真實的成分，我不深知，但是酸味是有一點。在雅虎的入口網站上鍵入「Data Mining」，立刻便找到一百五十多個網址。如果用 google 蒐尋，它在 0.34 秒裡就找到 1,260,000 項查詢結果。這是在方法論都還沒有成熟之前便已有大量的 commercial products 的情形。看起來真是商機無限。在另一方面，我卻在文獻上讀到：

**“..... while I appreciate the importance of data mining, in practice the profit it brings has turned out to be surprisingly limited in many key businesses.”
(Kann 2000)**

多半的 data bases 在建造時都別有目的，並不是設計來給大家 mining 的。上帝給我們以海，也許只是給我們以鹽。現在鹽的利潤有限，大家拼命去提鍊鈹，當然結果不會很好。

比較中肯的是 Hand et al. (2000) 的說法：

”Data mining is the process of seeking interesting or valuable information in large data bases”

方法

Demming (1943) 曾說：「蒐集數據的目的是為了行動」。這是在數據的蒐集並不便宜的時候所說的話。現在要反過來看：已經有了一大堆數據，當初只是為了蒐集而蒐集 —— 因為反正不貴，而且說不定某一天會有用 —— 問這裡

面可以提供怎樣的信息，能讓數據的所有人採取甚麼行動？

這意思仍然是要我們用心地去看這些數據。「用心看」有兩個角度：整體的 (global) 和局部 (local) 的。

Global view: 模型

自整體的角度來看一個 data set，是靠統計學裡面的抽樣就可以發揮得不錯的。抽樣的要點的 thinning —— 用一組較小的、容易處理的精選的 data 來反應整體。從這裡我們可以做一大堆傳統的建模工作。但這裡面最主要的觀念是抽樣。

例如 SAS 的 Enterprise Miner 中所建構出來的「table」，就是自全部的數據中所抽樣而得。他們引以為自豪的 biased sampling，在抽樣的理論裡行之有年，並不足奇。自抽樣的觀點來看，母體再大都不是問題 —— 它反而使得抽樣的理論更加簡單。何況在 data base 上抽樣，成本既低，又沒有我們最頭痛的 non-response 問題。這一部分，可以說是所有我們想做的理論和方法都十分成熟，只要找一群還不錯的人，將這些已知的事物組合包裝就好。在 table 上做傳統分析，最後的結果當然是一個可以用來描繪所有數據之間的關係的模型。這是傳統的計量方式：模型是所有資料的摘要。

Local view: Pattern detection

自局部的角度來看 DM，到目前為止，主要的目的是 pattern detection。這和我們常聽到的 pattern recognition 頗有不同。用雷達找飛機的工作，算是後者 —— 我們知道飛機是我們要找的對象，(這叫 supervised learning)。但是在 DM 中，多半我們並不知道要找甚麼 (這叫 unsupervised learning)。在技術上，這也不能靠抽樣來做。而主要的想法之一，是靠「算則 (algorithms)」而不是靠統計。Pattern 是指數據的局部結構，因此在這一部分，DM 強調的反而是算則。這當然並不奇怪，數據一多，用人力是不能做的，我們只有靠明確的指令讓電腦一個一個幫我們找。但我們全然可以不靠統計嗎？也不盡然。例如取訓練樣本 (training sample)，是很可以把實驗設計的想法放進去的。雖然，我們不見得要靠傳統的、精雕細鑿的抽樣設計。

工具

知名的軟體，像 SAS，SPSS，SPLUS，STATISTICA，其中多包含了一些關於 DM 的 functions 或者 procedures。一般還是離不了

- regression (predictive model)
- classification (supervised or non-supervised)
- decision tree
- association analysis (rule extraction)
- time series (for trend)
- machine learning (recursive partitioning)
- graphic methods (data visualization)
- neural network

等這一類的工具。這些工具或者算則，並不是每一種都有證明的。雖然，很多都有相當大規模的電算模擬。如果再配上 boosting, bagging 等花招，或者更當紅的 supporting vector machine，就足以像一個專家了。

應用

DM 主要的應用好像都在商業。而在商業裡仍然是以顧客的行為分析為主。下面的列舉，主要來自一個 IBM 的網站 (<http://www.almaden.ibm.com/cs/quest/>)：

Identify and extract high-value business intelligence from enterprise data assets, including :

High-volume transaction data generated by point-of-sale, ATM, credit card, call center, or e-commerce applications, Web navigation patterns.

最後那一點是我加上去的。不論怎麼看，這裡的商業氣味是蠻濃的：

Analysts and business technologists are empowered to discover patterns which might otherwise be unobserved, across volumes of data they were simply not able to penetrate with other types of analytical tools. Intelligent Miner provides the fundamental technology and tools to support the mining process, as well as application services to enable development of customized applications.

不久之前我曾聽過 SAS 的 Enterprise Miner 的三小時的說明。那時候我只瞭解到抽樣建模的那一部分，因此我對 DM 的定義也有一點負面：它相當於「elementary statistical methods applied to huge data sets」。但那算是 global view，並不會進到 algorithms 的層面上。對於做理論的人來說，我承認我還看不到甚麼大的東西。也許是因為 DM 主要還是要靠去用。而用是不一定需要數學證明的。我讀到的 DM 論文中，提到所建議的 algorithm 時，都蠻保守地說那是 heuristic。

如果你要進入這一行，當務之急是去弄一兩個大一點的 data bases 來玩一玩。玩一下才會知道問題在哪裡，哪裡才有有意思的問題。大型的 data sets 是搞資訊的人士這許多年來產生出來的 cyber 怪獸。怎麼去對付它，並要它替我們做事，的確是蠻困難的。但是，你總不能假設數據是來自一個 flat file，就算你還不會用 OALP 處理一個 data warehouse，但是，用 SQL 指揮一個 relational data base 做足夠的運作總是起碼要會的工夫。不懂這些縮寫？呵呵，去學啦。你若是要買車，總要知道 ABS 是甚麼是罷？

DM 的許多招式，來自統計和機械學習 (machine learning) 這兩種學問。但用在 DM 的實務上，都不免有些荒腔走板。也許是多半靠 DM 吃飯的人，這兩門課都唸得都不夠深入。我們一力地想把 CDMS 組織起來，這裡面是有一點心意的。

Miners 是一個行業。在 ACM 之下有一個 ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)¹，每年舉辦一個 KDD Cup 比賽²，並有一個學術期刊：Data Mining and Knowledge Discovery³，已發行到第五卷。很多作者都有自己的網站，找資料是不難的，但是找有用的資料就難了。

參考資料

Berry, M. J. and Linoff, G. Data Mining Techniques for Marketing, Sales and Customer Support. Wiley, 1997.

Deming, W. E. (1943). Statistical Adjustment of Data, Wiley.

Friedman, J. "Data mining and statistics: What is the connection?", The 29th Symposium on the Interface, Houston, TX, May 1997.

Hand, D. J., Blunt, G. , Kelly, M. G. and Adams, N. M. Data mining for fun and profit. Statistical Science, v. 15, no. 2, 111-131.

Kann, W. (2000) Comment on "Data mining for fun and profit" , Statistical Science, v. 15, no. 2, 127-130.

後記：

這文章我今年曾在福華飯店講過，但時間又過了大半年，我又加了一些話進來。二者有何不同？我想說的是：現在我對 Machine learning 比較知道一點了。對於 DM，知道的還不夠。

¹ <http://www.acm.org/sigkdd/>

² <http://www.acm.org/sigs/sigkdd/kdd2000/>

³ <http://kapis.www.wkap.nl/jrnltoctoc.htm/1384-5810>

二、關於東華本《統計學》(2007.2.9)

February 9th, 2007

為何要寫？

作者之一在退休後較有功夫

覺得時下的統計學教本 —— 包括不少外國書 —— 其內容都和四十年前的沒有不同。

意思是說：裡面的該淘汰的內容，因為別的課本也還有，都不敢拿掉；該加進去的內容，因為別的書沒有，也沒有加進來。

目前多半的教本，都提到電腦計算。可是，講的都是用電腦算用手都算得出來的事物。電腦的用處，是拿來算不能用手算的。

所以，這本書所講的電腦統計計算，很多是你大概用手在一小時或一天算不出來的。自七零年代以還，統計學最大的進步便是有效地和現代的資訊技術接軌，怎麼可以不教？

這書的特點

任何一個學門，如果還能在現代的科技社會裡立足，必需有一套完整的思路。不止是統計學如此，其它如物理、化學、生物等等，莫不如此。

如果只是應付考試，最有效的辦法就是多做考古題。雖然等考試的效用一去（如考取了或者放棄不考了），過一陣就會忘掉。為了應付考試，我們也有一本和補習班名師合寫的統計學（鼎茂出版社：《統計學評論》）。

如果你將來還想用到統計（例如初等應用：作圖列表，交叉分析之類；或者高級應用：建構模型，解釋現象，從而得到實務上或科學上的意義），那麼，你最好用「讀全本」的方式來「唸懂」。

這本書給你的是初等的部分，並且也試著向高級的方向鋪路。因此在講道理的地方一點也不馬虎。如果讀者還想更進一步，這本書應該能給你相當多的知識，讓你可以出發。

三、把春波都釀作一江春酎(1999.2.2)

一

電話八點鐘來，經過衛星的聲音 有些上氣不接下氣，但是相當職業化。「我是某某報的文教記者，」她說：「能不能向你採訪一下？」

她有興趣採訪的東西，大概不是我想說的話。兩三個回合之後，她問：「你對於這個中央研究院統計研究所的期望怎樣？」

「五年後大約能和一個美國的一流半的大學的統計研所比一比。」我老實說。

她的反應應相當失望。不值得寫一篇隔洋專訪，她一定覺得。但是一個一流半的研究所豈是容易造成的？我也不肯修正我的估計。樂觀的話是比較容易說的，也比較容易被接受的。何況我已經相當樂觀。

自從我認真考慮來負責籌備中研院的統計所以來，第一件事便是買一套曾文正公全集來讀。曾國藩是一個你年紀愈大你才會愈佩服的人，也是一個非常腳踏實地人。這個人的頭腦雖然裝滿了四書五經，可是做起事來，可能比牛頓還清楚。

從曾國藩那裡學到一件事：湘軍的功業，是一群志同道合的人共同努力的結果。

二

民國六十年的時候，周元燊、謝聰智和我，合力在南港開了三門課：機率、測度論和統計，怕沒有人來，周先生拿出絕招：來聽講的人，發鐘點費，請吃午飯，還發給計程車票。

火是不是這樣點起來的，我不確定知道。但是我想我們的確影響了好幾個好學生，而他們也直接或間接地影響後面的人。但這個時期內，「統計」還強烈地意味著「學術」。來聽講的人，清一色是大學裡來的。我們也知道，國內生物統計方面有多人埋頭做了不少工作。但是對另外一支主力——政府方面的統計工作——我們還缺少較深刻的認識。

這之後，周元燊先生在南港研究院停留了幾年，獎勵後進，不遺餘力。刁錦寰先生數度返國，引進了時間序列上的一些有用的工具。在這段時間裡，統計逐漸從學院裡走出來，一點一滴地和工業、人文、醫療診斷、工商管理等有血有肉的東西結合起來。

大學裡的統計系或者相關的科系也漸漸增加起來。回國服務的留學生，也穩定地增加。另一方面，品質管制也逐漸茁壯。中國品質管制學會，在這些年裡貢獻相當了不起。

統計逐漸成為一種行業。五年前，在中央大學主辦的統計方法講習會裡，不止可以招到三百名聽眾，更難能可貴的是，只有一半的人員，算是學院方面來的。很有些想不到的機關或公司，派人員來參加，例如像國稅局、榮民總醫院這些單位。

我也許可以舉一些今日已在有計劃、有組織地進行統計方面工作的公私機構。但這並不是我的目的，我是說，國內的統計氣象，已經稍具規模。而這些成果，都是十幾年來統計界同仁辛苦耕耘的結果。

三

中國的學者，做論文的本領絕對是世界一流的，只要政策是真心鼓勵發表的，中國的學者總是可以發表得又多又好。

發表論文當然是中央研究院的工作。但我不是為這個回來的。

我覺得，狹義的統計，是指一套科學方法，廣義的統計，則是一種制度。

只有一個華佗，或少數幾個華佗，縱然手到病除，也是杯水車薪，無濟於事。

如果華佗只是教學生，他的學生還是教學生……這還不是制度。

制度是：怎樣挑選好青年學醫，建立醫學院的規章，制定醫生的最低標準，發給執照。建醫院，制定護士的標準，藥劑師的標準。藥品上市的檢驗規格……。只有在有了這些之後，你才看得見平均年齡在增加，嬰兒死亡率在減少，吐痰的人日少，腰桿挺直的人增多……。

四

日本人只學了一招統計裡的品質管制，但是他們不但能徹底吸收，而且還融合到大和民族的特有文化裡去，演化至今日相當有民族性的日本品管。日本人是感念戴明博士的，但那實在是日本人自己的功勞，點火的是他，但發光的不是。

我們不難寫出水準比日本高的統計論文。

但要我們排得出「雲門」那樣的舞，那樣既是現代的，又是中國的舞，我們還是有不少的路要走。

統計的語言，本來就是社會的語言。只不過它更精、更簡，更觸到痛處。

我們不只是看平均數，我們看到的是平均國民所得，我們不只看到常態分布，我們看到的是一個群體：它既是一條平滑優美的曲線，也是一群鮮蹦活跳的人；我們不只看到兩個平均數的差，我們看到的是一種藥比另一種藥有效，一種肥料配方比另一種能生產更多的米，我們造出的模型，不只是一套方程式，而是社會的一個具體的縮影。

我們不能無動於衷，因為統計是入世的科學。我們觀察、整理、估計、檢定、模擬、預測……我們的主要對象是這個社會的形形色色。我們看到的山是山，看到的水還是水。而且因為我們是用有情的心境來看的，所以看到的山，不免嫵媚；看到的水，不免清秀。

五

我問：我們還預備作些什麼？

有一個人，才做一個人的事。有兩個人，才做兩個人的事。問題是：合作，還是分作？

規模稍大的統計工作，都不是一個人能包下來的。

我想：我們必須學會了合作一件事。

中國統計的種子，已種下不少，有的已經發芽，有的長得很好，有的還有些艱苦。

我想，我們最需要合作的一件事是：努力澆水、施肥、除草除蟲。

國內並不是不需要統計方面的宗師大匠，可是我寧可看到兩百個博士、三千個碩士散處在各個崗位上。因為在有了這種陣容之後，這個社會裡所產生出來的重要資料，才會被消化、吸收，濾去誤差，並且被大聲疾呼地引起重視。

讓我們時常相聚：統計、作業研究、電腦品管、工商管理、計量科學……凡是需要和常態分布，期望值、標準差……打交道的人士。現代的社會是一個以數量度量的社會。讓我們把這些科學融合使用。讓我們在一起談，這樣我們才逐漸說互相能懂的術語，才互相同情彼此的困難，才能培養在各方面都懂一些的人才。讓我們抽時間走出自己的領域，到你的母校中學演講，去了解一下國內的白手起家的小工商業裡面臨困難……。

讓我們能都體認到統計及其相關的計量科學，在台灣的發展途徑，是經過科技再進入科學。這種科技是觀念上的，是由觀念反應到決策及政策上的。它的觀念不能夠直接地被反映在國民生產總額之中，但這種科技是一種思想方法和思想方式，它幫助整個社會的環節，讓它能有效地製造數字、冷靜地分析數

字。

讓我們互相合作，多造就這方面的人才，並儘量為他們製造就業機會。支援他們、鼓勵他們，做幾件有長遠影響的工作。

這些工作，都是科技，不是科學。可是我們並沒有統計科學，我們甚至於還沒有廣泛地推動了統計科技。這些工作，都不是那些理性的、好奇的、純粹的名家大師的工作。這些工作，都不易得到學術界的讚嘆。可是，如果沒有人肯做，讓我們負擔起來。先求通，再求好；先求廣，再求深。因為沒有農業，園藝只是點綴，沒有廣泛地統計科技做基礎，統計科學也只不過是三個標準差之外的逸點而已。

後記

我喜歡的文章是這樣的：在哲理裡夾著詩意，平淡裡閃耀著智慧的火花。字裡行間，跳動著的是勇者的氣度與仁者的心懷。當你掩卷時，你得到的感覺，是三分灑脫，三分無奈，三分情愁，更一分穠麗。

這篇文章在很不情願的近乎呼口號中結束，這不是我喜歡的格調。這一星期以來，我的案頭，不斷翻閱的是一卷詩選，一卷詞集，再一本劍河倒影。

這些剎車發揮了多少作用？我不知道。但是當你有話要說的時候，你就忍不住愈說愈大聲，愈來愈臉紅脖子粗。

我一再強調統計是入世的社會科學。而社會科學所揭櫫的，明白地說，只是格致修齊治平之道。統計及相關的計量科學，只是對社會現象的格致方法。

統計因為入世，所以，只有廣、才能深；而真可能夠使之廣泛生根的辦法是製造一大批數量科學方面的工作機會及一大批正好可以派上用處的人才。因為這樣才有機會將近代的這些格物致知的方法，廣泛而有效地注射到這個經濟結構的各個重要關節中去。

這種工作有的朋友已經默默地做了。十餘年來，四次返國。每一次統計界的朋友所給我的，不止是許多驚喜而已。他們的貢獻與成就，才是我可以為此樂觀地的發表意見的原因。

p.s., 此文作於一九八一年，人還在 New Jersey 的時候。原先是科學月刊的邀稿，但最後並沒有刊載。這並不是好文章，也不是「科月」的風格。但我平生也沒有寫出甚麼好文章，因此二十年來還不捨丟棄。我歸類它在「統計散文」一類，因為它不談定理，又沒有公式。

到了一九九九年，再讀時才發現其實我的想法，二十年來一直都沒有變。這也許有兩個原因：我其實相當堅持；或者我一直不曾進步。此文曾在「統計薪傳」一卷一期發表。

四、昔年種柳(1982)

一

今年（民國七十一年）七月一日，中央研究院正式成立了統計學籌備處。這件事情至少有兩種意義：第一：這是政府對於過去二十多年統計界同仁的貢獻的一種正式認可：統計學自理、工、農、商這些學院的羽翼下跳出來，開始與別的獨立科學如物理、化學、數學等分庭抗禮；第二：我們意識到國內的統計學方面的研究、發展、應用和推廣，還有許多路要走。我們應該研究什麼主題？發展那些系統？如何推展應用？怎麼將統計學所表現出來的近代的格物致知的方法，廣泛地深植到社會的思想形態裡？我們應該認真地開始回答這些問題，而統計所的成立，給我們一個機會。

我想藉這篇文章，介紹一下統計學，國內的統計學以及我們想做的工作。

統計是一種科學方法和思想方式，是一種入世的、積極的哲學態度，是一種進步社會中自然的現象。

二

狹義的統計學，是泛指一套圍繞著以數量為主的資料的科學的格物致知方法。這套方法的一個最大特徵是：統計學家深切地體認到誤差的存在，並積極的面對可能的誤差，而使得經過這套方法所導出的結論，其因誤差而產生的曖昧減少。

統計學的方法並不能無中生有，但它的確致力於儘量濾去誤差，而得到傳統方法所不能得到的結論。誤差如水，真象若石。水落，所以石出。如果水中原本無石，水落當然也仍然無石。

統計的方法之所以大行其道，是因為誤差的是近代生活的一部分。一切人類所蒐集的數量資料中，其不包含誤差者百不得一。社會愈進步，則所需要蒐集與分析的(包含著誤差的)資料也愈多。而我們愈嘗試的去以無誤差的概念去推演結論，所得的結論裡便愈充滿誤差。而統計方法，因為能正視誤差的存在，反而可以得到更合理的結論。

因為這個原故，我們巧妙的自台灣地區人中抽樣，以求得平均國民所得或失業率；我們自少數病人的追蹤記錄裡研究一種新藥的效力；我們用精心設計的品管圖來看生產線上的狀況；我們小心地分配肥料與農藥的成份，而來看那一種配方使稻米的產量最多。

這些都是統計方法：圍繞著包含了誤差的數字所作的種種精巧的努力。這些包括了蒐集資料以及分析資料。資料，是統計學的食糧，氧氣和水。資料，是現

代的文字，是工業社會的文字，是科學的文字。統計學只是幫你去認字、去寫字，並用這樣的文字寫作品。

而這些作品所表達的是另一度空間的現代社會。它非詩、非畫；但更精、更簡，更觸到痛處。因為它的目的是探測真象，而它的手段則是濾去誤差及噪音。一方面，它像一個楔而不捨的偵探，搜尋線索，避開兇手所佈的疑陣，抓住了問題的核心；另一方面，它也像一個嚴正法官，考核證據，一絲不苟，既要冤枉，也要不縱。

我們不能無動於衷，因為統計是入世的科學，所以統計工作者也心在紅塵。我們觀察、整理、估計、檢定、模擬、預測……我們主要的對象是這個社會的形形色色的一切可以用數量測度的事物。我們看到的山，是山；我們見到的水，是水。而且因為我們是用有情的眼光來看的，所以我們看到的山，不免嫵媚；見到的水，不免清秀。

三

國內目前的統計，可概分為人文、生物及數理三類。這三支主流，以一個開發中的國家而言，都算得上相當源遠流長。而其他計量科學的支脈，如計量經濟、工商管理、工業工程、品質管制及保證、自動控制……也並不是沒有人才。自經濟發展的過程來說，我們的社會已繁榮到必需以接近天文的數字來描述的程度。大到政策的制定，小到一種商品的行銷，都離開不了計量方法。然而這些以統計學為基礎的近代計量方法，(還有以電子計算為輔的近代系統)，並沒有成為社會裡中上級官員或經理人員的思想方式的一部分。國內在統計學及其相關科學方面，固然人才不足，但並不是基本的問題。國內的問題在理論與實務的不配合，以及統計的工作環境的缺乏。國內並非不能培養統計人才；只是不易培養立即能夠上陣的統計人才；即使培養了，也不易將他們置於易於發揮功效的位置。就算這兩者都做到，也因為缺乏適當的環境，這些人才的努力，皆不免流於事半功倍。因此，我們有近五十年的政府統計資料，但其在民間的威信，尚待建立；我們的生物統計源遠流長，但至今尚推行不了一套完善的藥物管制制度，我們有將近三十名統計博士。但多散處各大專院校。論文雖然時有佳作，但對整個的統計甚至於計量科學的發展，並沒有發揮出風起雲湧的作用。

但是我們不必氣餒。我們自己不肯一顧的政府統計資料，有外國人在分析檢定，以期孤立出我國發展迅速的原因；十餘年來，品管工作者默默的建立起類似日本的品管制度；政府機構及民間企業裡也漸漸地增加了計量分析方面的人員。而在國外，不論在怎樣的統計學會的集會裡，都可找的到中國人在發熱發光。

這樣的成績固然堪差幸慰，但是如果我們看一看統計學及統計工作者在別的國家的經濟發展裡所作的貢獻，我們還有許多路要走。如果我們不能將近代的計量方法，注射到我們國家的經濟結構的重要關節；為果我們不能影響這個社會

的中堅份子，讓他們以客觀的、計量的、容忍並適度控制誤差的態度和方法來處理現代的企業經營，那麼我們即使有一二位名師大匠，五六十名有佳作的研究人員，我們仍然在實務上趕不上日本，理論上比不上印度，更遑論英美等國家矣。

中國的統計，並不是不需要名師大匠。中國也並不是沒有相當傑出的名師大匠。但我們如果向這個方向走，我想統計研究所在十五年內，還趕不上史丹福大學的統計系。中國的統計學家並不是寫不出高深的論文，但是統計學本身，在這個社會裡還沒有堅實的基礎，那些高高在上的論文，有若用極度的平衡技巧的支持的高空走鋼絲的表演者。表演雖然精彩，掌聲也未見稀疏，但是我們知道每年畢業的新的計量科學方面的碩士，有相當困難去找待遇優厚的工作；更不論那數百名那些相關科系畢業的大學生了。只有在這些計量分析人才，在國內成為各界爭相邀聘的對象時，我們才可以勉強說統計學的基礎，初初建立。

四

遠在民國十八年的時候，所謂的甘莫爾設計委員會，便建議國民政府於其下之主計總監部。這項建設經過立院會議之後，主計總監部的名稱換成了主計處，直屬國民政府，以下設歲計、會計、統計三局，分掌全國歲計、會計、統計事務。這是我國正式有統計制度的開始。其後主計處在民國三十七年改隸行政院，名稱曾一度稱為主計部，旋即改回。五十三年以來，內憂外患、不絕如縷，然而國府的統計工作如財務統計，基本國勢調查，國民所得統計、物價勞工統計、公務統計、公務人員及其工作之統計等，均在超然統計制度之下，維持至今。我國在民國二十一年便訂定了統計法(目前的三法修正，為民國六十一年五月二十六日公佈施行)。在各國政府所作的統計資料蒐集工作而言，我國是走在相當前面的。據六十三年二月全計處編印的「各項統計方法專輯」，目前我國至少有氣象、人文……到社會衛生，其他公務等十五項，九十六目統計資料。

我們不能說政府沒有做統計工作，我們只能說我們沒有好好的利用這些寶貴資料。

在學術界，我國的統計學者也頗為傲人。台大的農藝研究所開始用電子計算機作逐次迴歸分析的時候，國外所謂的「統計程式庫」也不過在萌芽階段，並沒有如今日一般地風起雲湧。好幾個農業實驗所的實驗設計，也為國內的品種改良做了註腳。

相對地說，我國的數理統計起步的反而相當的晚。

五

我們不要把自己分得太細太嚴：這是統計，那不是統計，而另一個分明便是作業研究……。如果我們這樣分，那麼每一個領域便的非常窄，我們不能自前人得到什麼幫助，也不能有效的幫助別人。

所有的近代計量科學，都和統計有關。我們無意併吞他們，應該如對待芳鄰一樣：我們借給他們油鹽，也希望有時候能自他們那裡去借柴米；我們鼓勵孩子們和家的來往，而希望有一天或能發展進一步關係……

我們也不要將理論和實務分得太嚴。理論如花，實用如果。不會開刀的醫生，和不會分析資料的統計學者一樣。花和果互相關連。有的果、種子發不成芽；有的花，結不了果。但是花和果都一試、再試。這是一個機率的世界，沒有失敗，便也不會有成功。如果我們在花和果之間，畫一條界線；如果我們在統計的理論與實務中間劃一條鴻溝，那麼花雖然香，總離不了溫室；理論雖佳，也了無意義。

讓我們去試著了解並尊敬到目前為止國內的計量科學所做的工作。他們的門，也許正鎖，但我們何嘗去用心敲過？他們的園地，也許貧瘠，但我們何嘗去用心幫著灌溉？當我們有百萬大軍的時候，再去分海、陸、空；再去分軍、師、旅、團。而現在，我們規模粗備，雖然望盡天涯路，卻幸而不是高樓獨上。我們應該試著去做一些大規模的團隊工程：不論是統計的理論還是實務，我們應該學會以二十個人合做一個計畫、三四個人合作一篇論文。我們還不是趙雲關羽，單槍匹馬，在千軍萬馬裡尚殺不出重圍。而在國外，比我們強得多的研究發展單位，用比我們多的人力物力，做比我們野心小得多的研究計劃。

後記

這是 1982 年的舊作，今日讀來，心情猶在。國內的統計研究工作，十八年來成長了十二倍，但學者們「單兵攻擊」的毛病依舊。

標題是後來加上的。取自「昔年種柳，依依漢南；今看搖落，萋愴江潭。木猶如此，人何以堪。」

1999 又記

昨日在過紅綠燈的時候碰到國偉。他問：有沒有文章？我是有幾篇雜文，而且有些還沒有發表。看看罷。 2000/8/10

五、中文學術論文的規格芻議(數學傳播,十六卷二期)

趙民德¹

中央研究院統計科學研究所

摘要

本文淺論以中文寫作數理類論文時的一般規格，以作為撰稿者的參考。本文只談浮面的要求，雖不涉及論文好壞的問題，但也有限度地說明了學術論文的一些基本精神。對於初次撰寫學術論文的年青人，或有幫助。

關鍵詞：方程式，文獻調查，表格，規格，圖，論文。

美國數學會分類索引(1990)：主要 00A99，次要 00A22。

1. 緒論

論文的好壞，決定了學者的好壞。但不論是好論文或是壞論文，文章的格式，至少在出現在期刊上或當作正式學位論文提出的時候，也相當程度地表現這位學者認真的程度。本文的目的，是希望能說明我們所需要的格式，作為撰稿人的參考。此外，某些格式的要求，隱藏了學術研究的一些基本要求，我們也順便提出我所需要的格式，作為撰稿人的參考。

為了說明方便，本文的格式，就是照我所想像的標準格式來排印。

論文以節為單位。第一節例為緒論(introduction)，最後一節也許是結論。其中各節均以數字 1, 2, …等作為各節號。論文在本文之前，先是標題，後為作者及所屬機關，再附上全文所引申的關鍵詞(keywords)及美國數學會的分類索引編號(AMS subject classifications)，然後才是本文各節。本文之後是附錄及參考資料。表格及圖片，請在正文該出現的時候用括號指出。(表 2.1) 或 (圖 3.4) 表示他們的恰當位置。

2. 一致性

維持一致的格式是敬業的表現，而敬業精神是好論文的必要條件。一致性的意思是說，同一論文內部之格式相同。格式二字，可大可小，大到全文的結構，小到參考資料的詳細辦法，通通要求滿足一致性。例如 Biometrika，嚴格要求括號要以[{()}]的次序出現。對於參考資料方面，要求尚未正式在期刊上發表的不可列入，除非至少已被接受。基本上，作者自己還沒有撰寫完畢的其他作品，不應作為參考資料，更嚴格一點說，作者自己沒有讀懂的其他作品，也不應該作為參考資料。

¹ 本文並非研究論文，但這裡是感謝支助研究獎助單位的地方。本文原稿是中國統計學報的論文規格說明。

3. 嚴謹性

論文文字力求嚴謹。有把握時，就說有把握的話；沒把握時，就說沒把握的話。如果作者說：有下列幾種常用方法，那麼至少他要有證據知道這些方法的確常用，否則他只該說「有」下列幾種方法。省去「常用」二字，正是作者負責的表現。

4. 分節

論文以節為單位，代表主題不同的內容。節中如有必要，可再分次節(subsection)。次節以小數表之，如 2.3 或 4.5。正文中引用前後節數時，用節號 (§) 表之，例如見§4，或見§3.4。多數的節數用雙節號 (§§)，如見§§4-5，意指見第 4 節及第 5 節。

每節及次節均有小標題，例如

4.1. 這是§4.1 的標題

在這個標題之下，分段寫出§4.1 中的內容。

4.2. 這是§4.2 的標題

事實上，§4 中，如果只有§4.1 也不對，但若同時也有§§4.2-4.3，則較能表現節中分次節的意義。

5. 方程式

方程式原則上應儘量少用繁雜難以排版的形式。通常較難的公式，可改為二個較簡的公式。用「/」代替「÷」分開分子及分母，也有益處。用 $\exp(x)$ 代替 e^x ，也是一般的論文印刷習慣。方程式請用(4.3)，表示§4 中的第三個方程式。原則上，方程式不一定要編號，編號表示這是以前曾或未來將引用的式子。方程式的號示要連續。如果原稿中用了(4.1)，(4.2)，(4.3)，而在修正時取消了(4.2)，那麼請勿忘記將(4.3)升級成(4.2)。方程式的號示如果出現不連續的時候，表示原論文的作者用心程度不足。

方程式的編號，寫在右邊，例如

$$y = X\beta + \varepsilon \quad (5.1)$$

是一個標準記法。英文的"if"，在中文的方程式中比較顯得為難。例如

$$\begin{aligned} f(x) &= 1 \text{ if } 0 < x < 1; \\ &= 0 \text{ elsewhere.} \end{aligned} \quad (5.2)$$

若用中文寫，因為中文沒有倒裝句法，只能寫成：如 $0 < x < 1$ ，則 $f(x) = 1$ ；此外 $f(x) = 0$ 。如方程式太大必需用(5.2)的形式，則不如省去英文，記成

$$f(x) = \begin{cases} 1 & 0 < x < 1; \\ 0 & \text{其他情形。} \end{cases} \quad (5.3)$$

這樣雖仍有點勉強，但至少不夾雜符號之外的英文。

6. 定理

定理不妨分為引理(lemma)，定理(theorem)及系(corollary)等。通常，定理是主要的結果，引理是為證明定理而引入的輔助結果，而系則指主要定理之後所引申的結果（例如特殊化或技術上明顯但意義上仍存在的推廣）。我們不十分區分命題(proposition)與定理之不同。

定理，引理及系的編號，編號的原則和方程式相同。例如定理 3.2 表示§3 中的第二個定理。

作者當然希望被引用的是定理。但有時也有引理非常有名，但定理反而埋沒的情形。極有名的結果，可用縮寫，例如 SLLN, CLT。有人名的定理，請引用人名，例如「利用 Neyman-Pearson 引理…」。

如果這個定理還不十分有名，請給參考書目或論文。

7. 引用

引用(citation)表示承認被引用者之研究成果，除非是極有名的結果，如 CLT，請正確的引用。這意思是說，引用要恰到好處。某結果如確為某人所首創，那麼不要引用比他更後面的作者，例如提到乘積極限估計量(product limit estimator, PLE)，應該提 Kaplan 及 Meier(1958)，而不是 Lo(1985)，除非你所討論的內容，是 PLE 的漸近最優性。而在 1958 年到 1985 年之中，尚有一些重要的相關文獻，是否引用，端看作者的品味。引用時的人名如係外人請不要譯名，我們不覺得愛弗龍會比 Efron 更好。原則上，外文著作，請用原來的外文，例如 Chao and Lo(1985)，而中文著作請用中文，例如趙民德，田昕(1986)。同理，請直接寫「Markov 鏈」，而不必說「馬可夫鏈」。雖然，Markov 只是俄文的英文譯名，並非真正的原文。你若能找出原來的俄文，當然更顯得有學問。

8. 縮寫字

除非是極有名的縮寫字，例如 CLT, MLE，請在第一次正文中出現時附全名後再給縮寫，例如個人電腦(personal computer, PC)。在以此法正式介紹之後，便可自由在後文中以 PC 代替個人電腦了。

9. 名詞

我們並不堅持名詞的譯文必需統一。但請在同一篇文章中，同一名詞請用同一譯名，這也是一致性 (§2) 的堅持。當然，如期望值，標準差等人人皆知

的名詞，不必解釋及附原文。但讀者可能不明白的譯名，請在首次出現時附原文。

10. 圖及表

圖請附一個可即製版的圖，並附上圖上的說明(caption)。圖表如果不大，可直接在正文中出現；否則請給一個清楚的可製版的表（千萬不要只給我們一張色帶已淡的電腦報表），而將之放在論文之末。一個簡單的方法，是用雷射印表機，用你的判斷減少圖表的數量及大小。圖表的編號和方程式的編號，原則相同。太多的圖表，暗示作者沒有在多種資料裡擇取精要的能力。

11. 中文文法

論文是可讀的文章，讀中文論文的人，一樣也循著正常中文的邏輯思考來體會作者的所想表達的意念。因此論文中的中文應也求通順明晰。理工學者常有生吞活剝地硬譯英文的那種中文出現。應力求避免。例如中文沒有倒裝句法，因此像若且唯若這樣的句子讀起來就慘不忍睹，中文裡的「是」字，往往不是"verb to be"作者要自己分得清楚。中文也可以和數學符號混在一起，使數學的推導為明順。例如，令 $f(x) = 3x + 4$ ，則 $f(x + 3) = 3x + 13$ 。因此 $f^2(x + 3) = 9x^2 + 78x + 169$ 。這樣的寫法，比

$$\begin{aligned}f(x) &= 3x + 4, \\f(x + 3) &= 3x + 13, \\f^2(x + 3) &= 9x^2 + 78x + 169,\end{aligned}$$

要看起來像人性化的論文。

12. 感謝

感謝人的部份，請標在文尾結論之後，另開一段，但不要一節。感謝研究基金之支助部分，請在作者姓名上作*#等號碼，而在論文首頁之末，加線作足標說明之，如「本文受國科會部分支助」。請注意到，除非你的全部薪金來自該機構，否則只能感謝部份資助。這也是嚴謹的意思。

13. 附錄

附錄通常包括本文中提及，但未完全處理的細節（例如證明）部分。某些項目歸入附錄的目的，通常是為了使正文不致為技術上的細節打斷。因此，並不是所有的證明，都希望併入附錄，完全要看"文氣"而定。如有二、三個附錄，可用附錄 A，附錄 B 等。附錄中，公式、定理等皆可編號，但用(A.3)，(B.4)等。

14. 參考資料

參考資料放在文尾最後一部分。請以中文、日文、西文之次序，其中並以第一作者之筆劃或字母次序排列之，如有二篇文章，是同樣作者，則發表在先

者，先列之。

所有的作者，均需列舉。不能說「張三等」，因為別人辛苦作研究，就是希望有人引用，而被引用是學者頗看重的事，因此不可辜負別人。姓應在前面，以逗點分開，如"Smith, S.K."，二個人以上時，應寫作"Smith, S.K. and Wang, T.L."其後附年份。論文題目，如係討論論文(discussion paper)要另加"(with discussion)"。論文題目中，除第一字的第一字母及專有名詞首字及縮寫的全部均需大寫之外，均用小寫，請見本文之參考部分。至於專書之書目，則各字除介係詞及非第一字的冠詞之外，各字之第一字母，均用大寫。期刊的名稱、卷別、期號、頁碼均需註明；如係書籍，請附原發行書局名稱。翻版書請勿用在台灣的書局名稱，因為看起來實在很不雅。學術味道較薄的印刷物，如報紙、休閒刊物或通俗性科學刊物，應儘量避免。在學術論文引用讀者文摘，不是很有面子的事情。

為什麼要將參考資料寫得如此詳細，基本的理由是要讀者若有需求，他們有絕對的資訊去查出這些論文。這是學術論文的精神：無一字無來歷。

15. 結論

本文的主要討論，限在中文的規格，雖然間或也言及論文規格的精神。簡單的說，學術論文要嚴謹、誠實，具有一致性。任何結果，均應可獨立複製。我們的規格不過是這些精神的一些表現而已。

論文的格式精嚴，並不表示內容也好。但格式鬆散，前後不一致的論文，足見作者與編者都不用心，因此品質不好的機會頗大。劉備說「勿以善小而不為」，放在論文的寫作上，也很恰當。

感謝：本文承謝中天先生一再設計 $\text{CT}_\text{E}\text{X}$ 之規格，又承王惠青女士之鍵入，在此一併致謝。

附錄 文獻調查之意義

徐賢修先生曾說，文獻調查也是研究的一部分。為什麼？因為研究者也，是指研究出新的結果而言的。為什麼是新的？因為舊的文獻上都沒有！如果作者不去查明，他就不曉得他所得到的結果是不是別人幾年前已作出來的。作為研究論文的作者，他有義務去查明此事。因為只有新的結果，才能用他的名字發表。發表後如果引起重視，這位作者才會漸漸有學術地位。

時下的索引相當的多，例如 Current Index to Statistics 及 Comp. Math. 的 Citation Index，都很方便。較好的圖書館中，都有 Mathematical Review。因此現代的研究者，更沒有理由不去做文獻調查的工作。有了以上工具書的幫忙，通常兩、三天到一個星期，足夠相當程度地查出目前所做的工作，是否已做過。當然，偶然也因人世間還有其他尚未發表的文件，作者不知道他人的研究報告已做出。這種誠實的獨立分別做出，是可以容許的。

文獻調查的另一意義為給該得到認可的論文予以應得的認可，因此作者在
做這項工作時，要知道把功勞(credit)給該給的人。這可使得做出劃時代的重要
結果的人，得到更多學界的認同，也間接地鼓勵了學術研究。

文獻調查的第三種意義為它所隱含的推理。例如某論文在 1989 年寫出，關
於文獻，只引用到 1985 年，這時，作者應負的責任是：1986,87,88,89 都沒有任
何人做出同樣的結果。如果有人做而作者不知道，那麼他至少要忍受學問不深
的評語。這件事發現之後，他的論文也會變得全部或部分沒有意義——視雷同
的多寡而定。

如果必需的話，千萬引用別人的論文。否則，"抄襲"兩個字在學術界非常
難聽。不論怎麼說，寫論文是學者的風雨名山之業。學者，有學問之人也。他
們必需背負起文獻調查的十字架。

參考資料舉例

趙民德，田昕(1986). 以 EM 對單一設限數據求 MLE 方法之商榷. 中國統計學
報，v. 24, no. 4, 11100-19.

Chao, M.T. and Lo, S. H. (1985). A bootstrap method for finite population. Sankhya
A, v. 47, pt. 3, 399-405.

Lo, S. H. (1985). The product limit estimator and the bootstrap: some asymptotic
representations. Prob. Theory. Related Fields, 71, 455-65.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete
observations. J. Amer. Statist. Asso., v. 53, 457-81.

A Style Guide to Some Academic Papers

Min-Te Chao

Institute of Statistical Science
Academia Sinica

ABSTRACT

This is a style guide to papers submitted to the Journal of the Chinese Statistical
Association. We only discuss superficial requirement, and do not touch the quality of
papers.

Keywords and phrases. equation, figure, literature search, manuscript, style, table.

AMS 1990 subject classifications. Primary 00A99; Secondary 00A22.

六、統計已死，統計萬歲！(1999.7.15)

八十八年七月十五日
在南區統計研討會 陸軍官校講

人生有很多無奈事，其中之一就是被推到臺前講自己其實並不十分夠格的講演。尤其是這次南區統計研討會和張文豹先生基金會的安排，是中國統計史的第一次。我覺得，若是講統計學，就該從它的源起、傳承、今日的國際現狀和臺灣、中國統計學門的處境，一直談到未來的發展及方向。但這樣架構實在太大了，同時，統計這一行，不止是在臺灣，即使以世界的角度來看，也不是甚麼人多勢眾的學門。

做統計的人，最應該瞭解，不論自那一種統計數字來看，臺灣都是一個小國。那麼，務實地看，一個小學門，小行業，在一個小國家裡，應該扮演甚麼角色？

我是有一點故意地訂今天的講題。其實，以一個人口二千萬多一點的小國而言，臺灣的統計工作者，不論是官方還是學界，都相當成功地負擔起他們的任務。例如表一所示的，是「經濟學人」(Economists) 在 1995 年 3 月所做的對新興國家的官方統計數字的品質的評比。

表一：經濟學人對新興國家統計數字的評比

	Timeliness	Availability of quarterly GDP	Overall quality of statistics
Chile	4	Yes	A
Hong Kong	21	Yes	A
Israel	6	Yes	A
Singapore	17	Yes	A
South Korea	2	Yes	A
Taiwan	1	Yes	A
South Africa	8	Yes	A-
Czech Rep	5	Yes	B+
Argentina	12	No	B
Greece	14	No	B
Hungary	10	No	B
Mexico	6	Yes	B
Poland	9	No	B
Portugal	16	No	B
Turkey	10	Yes	B

Brazil	14	Yes	B-
India	17	No	B-
Malaysia	19	Yes	B-
Philippines	21	Yes	B-
Thailand	10	No	B-
Indonesia	19	No	C+
China	10	Yes	C
Russia	21	Yes	C
Venezuela	24	no	C

Source: Economists March 4, 1995.

表二則取自 Canadian Journal of Statistics 在 1997 的一篇論文，它做了一個相當廣泛的蒐尋和比較，我所取的只是其中的幾個較全面性的比較排行，講的是對於各國統計研究論文數量：

表二：Top 25 countries for gross national publication of statistical research.

Rank	Country	PAG*	PAG	ART*	ART
1	USA	138544	77974	9484	5493
2	Canada	20933	11805	1588	925
3	UK	16296	9949	1178	749
4	Australia	13891	7782	991	578
5	Germany	8868	5919	622	419
6	Netherlands	6939	4063	468	282
7	Japan	6813	4570	554	379
8	France	5034	2595	351	177
9	India	4252	2322	456	250
10	Denmark	3416	2105	211	135
11	Israel	2589	1450	190	109
12	Taiwan	2489	1433	182	110
13	Norway	2014	1303	126	81
14	Spain	1952	931	138	70
15	Belgium	1947	932	116	55
16	Poland	1784	1050	163	98
17	Sweden	1723	1128	121	83
18	Italy	1428	823	105	62
19	New Zealand	1401	929	115	73
20	Switzerland	1358	754	91	54

21	Finland	1340	702	96	51
22	China	1287	709	101	58
23	Brazil	967	448	72	35
24	Greece	945	572	76	45
25		841	548	787	53

Source: Genest, C. (1997), Can. J. of Statist., v. 25, 427-443.

至於表三，我也藉這個難得的機會報告一下敝所的期刊「中華統計學誌」目前以 impact factor 為依歸的排行情形。

表三：統計期刊的排行

Rank	Journal Abbreviation	Impact Factor
1	STAT SCI	2.467
2	T ROY STAT SOC B	2.141
3	J AM STAT ASSOC	1.851
4	ECONOMETRICA	1.712
5	J ROY STAT SOC A STA	1.556
6	BIOMETRIKA	1.446
7	SOLID STATE NUCL MAG	1.370
8	J CHEMOMETR	1.174
9	TECHNOMETRICS	1.103
10	STAT MED	1.003
11	INT STAT REV	0.973
12	BIOMETRICS	0.938
13	STOCH HYDROL HYDRAUL	0.909
14	APPL STAT-J ROY ST C	0.852
15	OXFORD B ECON STAT	0.781
16	ANN STAT	0.729
17	BRIT J MATH STAT PSY	0.667
18	AM STAT	0.633
19	PROBAB THEORY REL	0.624
20	J QUAL TECHNOL	0.608
21	ANN PROBAB STAT SINICA	0.602
22	STAT SINICA	0.538
23	MULTIVAR BEHAV RES	0.491
24	ADV APPL PROBAB	0.479
25	INT J GAME THEORY	0.473
26	J APPL STAT	0.448
27	STOCH PROC APPL	0.445
28	INSUR MATH ECON	0.412

29	J THEOR PROBAB	0.398
30	SCAND J STAT	0.396
31	J APPL PROBAB	0.387
32	ANN I H POINCARÉ-PR	0.379
33	ANN I STAT MATH	0.368
34	FUZZY SET SYST	0.346
35	QUAL QUANT	0.314
36	J MULTINARIATE ANAL	0.301
37	STATISTICS	0.299
38	COMPUT STAT DATA AN	0.275
39	J STAT PLAN INFER	0.263
40	BIOMETRICAL J	0.260
41	CAN J STAT	0.260
42	COMPUTATION STAT	0.255
43	J AUST MATH SOC A	0.250
44	STAT COMPUT	0.250
45	STOCH ANAL APPL	0.229
46	STAT PROBABIL	0.203
47	COMMUN STAT-THEOR M	0.194
48	S AFR STAT J	0.174
49	STAT PAP	0.164
50	J SATE COMPUT SIM	0.156
51	COMMUN STAT-SIMUL C	0.142
52	UTILITAS MATHEMATICA	0.123
53	THEOR PROBAB APPL+	0.063

Source: Journal Citation Report, Winter 1998.

相對地說，我們統計界的同仁的表現，不論是官方還是學界，都算是十分搶眼。但這些都是我所知道的好的、不是由臺灣的人或者單位自己調查出的數據。我們當然也可以找一些並不漂亮的數據來顯示統計這一行較差的一面。例如如果我們去看一下歷年統計科系的聯招錄取分，並和最熱門的一些科系作比較的話，我們也立刻可以看得出來，對大學部而言，統計科系，絕不是目前有志青年心目中的有出路和有出息的選擇。對研究所而言，這個情形也許略好，今年的各統計研究所招生，來考的人數大約在 100-200 之間，相較於一些聽起來好像較有出路的科系來說，是差了不少。至於就業的狀況，我手邊並沒有確切而全面的數字。但我想在座的很多人都知道，一個需要博士學位的教職有十到二十個人同時競爭是頗為平常的事情。

我們當然相信「數字會說話」。但還有一些數字沒有說的：

國科會在兩年前成立了「理論中心」，目前的預算大約是 3,600 萬，但其中並不包括機率和統計。

這個信息，和 1982 年中研院成立統計所籌備處所暗示的，是截然相反的。當年，有一些有影響力的學者，覺得統計是一門應該在國內發展的學門，

因此，他們才建議成立統計所，以備這樣的用途。而二十年後，另有一批有影響力的學者，反過來認為統計學不算甚麼數學，甚至於機率論也不算甚麼數學。或者，不算好數學；至少，是不值得國家特別拿一筆錢來支持的數學。

對於這樣的訊息，雖然我的朋友裡是有一部分表現了不滿與反彈，我倒是沒有特別的反應。一部分的理由當然是我的遲鈍，主要的道理，卻是我基本上相當程度地同意以上的關於統計的推論。統計之不算數學，其道理是和會計、財務、電機、不算數學是一樣的。理論中心不包括統計在內，固然可以解釋成「不值得國家特別拿一筆錢來支持的數學」。但我寧可將統計看成「不需要特別被保護的學門」。為甚麼要被特別保護？因為統計不是熊貓，不是雲豹。統計是以數據作為食物的動物。而現代的數據，已多得到了可怕的程度。老鼠若是掉到了米倉裡還會餓死，當然是老鼠的錯。統計這個行業在現代這個充滿數據的社會裡為何還要和一些不食人間煙火的數學家爭那區區的一塊小餅？我想，我們不要把看己這個行業看得太小了。

當然我們會去用一些數學，也許還用一些不是那些非常基礎的數學。但是統計並不是以做定理為主要目的的。統計本來不算是數學，當然更談不上是好數學。如果我們總是執著在定理的證明與推廣，爭執於某些一般條件 (regularity conditions) 是否較弱，收斂的速率誰比較快，這一類的遊戲是永遠玩不完的。我雖一向不認為自己是數學家，也看得出來這一類的研究，其實不能算甚麼好數學。多做、少做，也許會對國科會每年在計算 SCI 論文篇數時會為我國增加幾個百分點，但其它的影響，我就看不出來了。

即使是這樣型態的工作，臺灣學者的影響又在那裡呢？表四列舉的是 1995 到 1998 裡臺灣的學者在 *Annals of Statistics* 的所有論文：

表四：1995-98 之間國人在 *Annals of Statistics* 的論文一覽

作者	topics
95	
胡殿中	Density estimation
劉惠美	More powerful test
姚怡慶	Testing AR models
趙民德	Truncation data
陳 宏	Semi-parametric GLIM
任眉眉	Confidence sets in arbitrary dimensions
96	

何淮中	Long-range dependence
黃榮臣	Asym. honest sets
謝復興	EPA
97	
鄭明燕	Bandwidth selection
趙晨慶，梁文騏	CLT, permutation stat.
鄭明燕	Boundary effect

前面我就說過，我們是一個小國。在國際市場上，小國只能在別人的遊戲規則下玩遊戲。即使在這樣的條件之下，我們的高手也偶然會投進幾個球。我們當然要為他們的得分喝采。事實上，國科會在這方面的表現並不算差。明眼的人可以看出來，這張表裡，頗有一些傑出獎的得主。我們雖然不能說這個表中的人物代表了全部這個島上的菁英，但這些人物，的確是略為走在前面的。

這些人，努力地匹馬單鎗殺入敵陣，可是，他們的大部隊又在那裡？他們是不是會太孤軍深入了，以至於後援不繼，被別人切斷了後勤補給？

我很久之前就說過，一個像樣的學門或者行業，需要一個漂亮而穩定的金字塔型的結構，這樣才能長久地讓頂上發光。

那麼，我們來問一問，誰來讀統計？

早在八十一年，我在某一次統計年會裡就說過下面的一段話：

「正面來說，華裔學者在世界的統計學術舞臺上佔有顯著的地位；負面來說，這樣金字塔尖端，並不是建築在國內的廣大基礎上的。

「在六零至八零年代，這一隻隊伍，源源不斷地被臺灣及香港出身學有專精，並勇往無前的數學系畢業生補充壯大，在國際學術舞臺上，的確扮演了一些角色。但隨著數學系的兵源日稀，經濟發展良好之後的學子的眼光更放在企管、財務金融等更入世的學門，我們或多或少地覺得這方面的盛況，在可見的將來，必會讓出一些地方給大陸的留美學者。早年的成功，建築在早年的一些機運和條件。當這些條件不再或較難存在時，那麼這些成績便會減色。這是一切不在本土生根文化的宿命。」

這個問題有它的歷史及時代的背景。早期，數學系出身的學子，投身於統計這個行業，是國人能在國際舞臺上佔一席地的主要條件。但這個優勢已不復存在。例如哥倫比亞大學的統計系，這幾年來的 entry level 工作的申請人裡，並沒有看到臺灣的留學生。但大陸同學的申請案，近兩年每年都在六十件以上。臺灣在美國沒有新的 Ph.D.產生嗎？在座的諸位，不乏各系所負責延聘新人

的教授，所知當比我還要更第一手。我想，也許我們應該承認，目前臺灣的留學生，至少在統計這一行，已逐漸在頂尖研究單位裡和不能大陸的留美學生競爭。

讓我們看一下我們的大部隊在那裡，而他們在接受怎樣的培養和訓練。下面的表，是我托朋友蒐集的國內需要考數理統計的研究所。

表五：要考『統計學』的研究所

台大	經研所、圖館所、工工所、商研所、資管所、地理所、土木所（交通工程組）、國企所、農經所、會研所、財金所、流病所、農藝所、植病所
政大	企研所、財管所、資管所、科技管理所、經濟所、國貿所、圖書資訊所、地政所、保險所、金融所、統研所
交大	資管所、工業工程與管理所、科技管理所、統研所
臺科大	管理技術所(企管組資管組工業管理組)、營建工程所(管理組)
成大	企研所、國企所、會研所、統研所
中央	企研所、財管所、產經所、人力資源管理所、土木所、資管所、工業管理所、統研所
中山	資管所、財金所、公事所、海環所、大陸研究所
中正	企研所、財務金融所、國際經濟所、資管所、數統所
中興	企研所、經濟所、都市計劃所、農推所、統研所
暨南	國企所、經研所、公共行政與政策所
雲科大	企研所、資管所、工業管理所、環境與安全所
東華	企研所、國企所
屏科大	資管所
輔大	管研所、經研所、應統所
元智	管研所、工業工程所
中原	企研所
逢甲	企研所、工工所、統計與精算所、保研所、建築與都市計劃所、土地管理所、電機所(機率與統計)、教育研究所(教育統計學)
台中師院	測驗統計所

這個表比我所能想像的要廣。我相當驚訝地發現，國內要考數理統計的研究所有那麼多。我的一位補習界的名師朋友告誡我，補習統計學以備考研究所的市場，每年是 15,000 人。

這個數字也比我能想像的多。而且，這 15,000 人，大部分不會被在座的各位教到。這些人一生中可能這一段時間裡是最用心讀統計的。將來他們在統計分析上的功夫和招式，也必然地深受這一段教育的影響。他們讀到的是怎麼樣的統計概念？

例一：一本補習班的講義如是說：

「獨立和互斥沒有強烈關係。」

例二：下面是交大 86 年「科管」的考題：

How is the formula for the variance affected if a body of data is a sample rather than a population?

- a. The variance is used for population data and the standard deviation is used for sample data.
- b. The squared deviations of the measurements from the mean are divided by $(n-1)$ for the sample and divided by N for a population.
- c. The squared deviations are used for a sample and the square roots are used for the population data.
- d. There is no difference in the formulas.
- e. None of the above

近年來，我偶然會做一點「箋註」的工作。對於「箋註」這兩個字，國語日報辭典的解釋是：註解古書。做學問的方法之一，就是重新詮釋古人的精意。例如朱熹的主要成就，來自他對孔孟之學的註解。而我們這一行裡，B. Efron 有好幾篇非常有名的論文，究其實際，也只是用心地再一次解釋 R. A. Fisher。我的註釋工作，目標要小得多。我只想去弄懂並說明一些應該可以弄懂，但目前國內的資料裡還不是很清楚的東西。而我的對像是在 elementary level。當然，這些基本上都是自己已沒有極好的主意，做不出有新意的結果，並在讀別人的東西或讚嘆或不爽時的閑話。例如我一直弄不懂品質界的 α -系統。直到有一天弄到了一冊 Motorola 的第二代的版本，讀完之後，啞然失笑，因此就寫了一篇「某公司講義箋註」。我對例二的註解是：

「這一題，英文不知道他想說甚麼，統計也不知道他想問甚麼。遇到這樣的考題，你就交卷出來算了。你已經浪費了報名費，還想再浪費兩年學這樣的東西嗎？」

但是，這一批人卻可能是日後大量使用統計方法的大軍。有的時候，我會有一點覺得，我們做的定理好不好，問題其實不大。

因為反正沒有甚麼影響。

但這些人所學到的統計方法是對是錯，是精是粗，反而是真的問題。因為這些人可能是負起國內大多數計量工作的實際操作任務的一群。對於前面我所提及的已殺入敵陣的菁英族群而言，這些人能提供的後勤支援就十分可慮了。我的意思是說，我們有相當可以和國外相比較的研究工作人員，但很可能我們造不出大的系統。因為做基本工的基層，不止沒有受到足夠的訓練，有時還可能被誤導。統計這個行業，尚沒有成氣候。用更明白的話來說：那就是，統計本身，尚不能提供大量的工作機會。

甚麼是統計這個行業？我不覺得我可以明白定義，但可以舉一二個例子。

例三：Barra US Multi-factor Equity Model 的 E2 模型

基本上，這是推廣 Markowitz 的一系列關於證券組合的論文所製造的模型。經由 Professor Barr Rosenberg 和 Professor Vinay Marathe 的研發，於七零年代早期，在 Berkeley 所建立的。

證券組合的理論是說，任一證券的報酬率，都可視為一個隨機變數。因此它就有相應的 moments。平均值就相當於平均報酬，而變方則可視為該證券的

風險。給定 n 種證券，則我們便會有 mean vector 及 covariance matrix。若我們知道這參數的值，則利用 quadratic programming 的方法，就可計算出最佳的證券投資組合。這在 n 大的時候，同時估計這樣多參數，就成為一個問題。簡單的方案是引入結構，以降低參數的數目。可所用的方法，則是 factor analysis。Barra 模型，最後可以降到 68 個 factors。

如何建構這樣的模型當然是苦工和商業機密。W. Sharpe et al. (1995) 在 “Investment” 一書中，用了很大的篇幅來描述這個模型。但主要的工作可分為五個步驟：

- Data collection and checking.
- Factor selection.
- Creation of composite factors.
- Estimation of factor returns and the factor variance-covariance matrix.
- Model testing.

我只能在這裡說，這些都是統計分析的基本動作。以我國的實力來說，主計處的一位第三或第四局的科長，如果有充分的人力和資料的支援，就應該做得出來。

以上的道理在座的人都懂，但 Rosenberg 卻看得出商機並加以實現。他開了一家叫做 BARRA 的公司，目前的年收益超過美金 4,000 萬。

這一年來，我頗喜歡告訴別人這個故事。這是統計的行業，但你若不主動出擊，就會變成別人的行業。

別的例子諸如 ETS, A. C. Nielsen, Inc.。前者做測驗，後者做市考調查。但他們的基本技術，全是統計學。

當然，你可以說，他們還有別的東西。Barra Model 的後面是證券財務管理，ETS 後面的教育測試，而 A. C. Nielson 的後面是市場學。話說到這裡，其實意思已十分明白：你不能為工具而發展工具。工具的目的是使用，只有在極度使用之後尚嫌不足的時候，發展特殊工具才會帶動實質成長。

用最底層的測量來看，那就是說：統計及相關科系的畢業生的就業情形有沒有實質成長？

前面說過，統計是以數據為食物的動物。

統計的本業是消化數據，並產生有營養的結果。它的本質，和母牛相差不多：

DATA-----> STATISTICS ----->INFORMATION

GRASS----->COW-----> MILK

當數據型式改變的時候，我們就應該發明新的方法，以消化新的數據型式。驗諸統計方法的發展，我們發現在數據少的時候、在數據不完全的時候、在數據很多的時候、以及在數據極多的時候，統計全都有不同的方法、角度、邏輯、甚至哲學態度去面對。目前，好像並沒有甚麼統計人去做 data mining 或者 machine learning。但這至少是一些如何去面對鉅量數據的問題。我們知道，任何 model 大概都是不對的。因此，利用概度函數及有限幾個參數的那一套傳統方式，在現代的數據又多又容易得到的時候，就會面臨嚴峻的考驗。

我們看統計學的發展史，基本上是在看它因數據的型式或來源的改變而產生的對策。當氣候已然改變，或者我們已預見改變的時候，例如電子及網路已使得極大量的數據可以很容易地取得，如何去調適統計的技術能力，其實是攸關這個行業生存的問題。

統計學是工具。工具必需要有使用的空間及場合，才能見到真正的效益。

「神之於形，猶如利之於刀。未聞刀沒而利存，豈聞形亡而神在哉！」

上面這段話，幾十年前曾對年青的胡適產生極大的影響。「鋒利」是磨刀的結果，但一味只是磨刀，是看不到結果的。在三船敏郎所主演的「宮本武藏」裡，有一段場景：宮本拿武士刀給一個磨刀師去磨，磨刀師充分地表現了磨刀這一行的尊嚴，他對武藏說：

「我磨的是武士的靈魂」

話是說得十分高級，但磨完的刀仍然是給武藏拿去做決鬥用。這個電影，如果只是演到磨刀完成就停止，就不會有票房了。

統計有沒有死？會不會萬歲？

只要有米倉，就會有老鼠；只要有數據，就會發展處理數據的方法。但是不是叫做統計學、或者叫做 computer science 的 data mining，就要看這一代的統計人如何因應變局。

那麼，甚麼是統計人？在座各位的認知，我的認知，以及一般人的認知，恐怕都是有些差異的。多年以前，我讀到下面一段有趣的文章：

“You haven’t told me yet,” said Lady Nuttal, “what is your fiance does for a living.”

“He is a statistician,” replied Lamia, with an annoying sense of being on the defensive.

Lady Nuttal was obviously taken aback. It had not occurred to her that statisticians entered into normal social relationships. The species, she would have surmised, was perpetuated in some collateral manner, like mules.

“But Aunt Sara, its a very interesting profession,” said Lamia warmly.

“I don’t doubt it,” said her aunt, who obviously doubted it very much.

“To express anything important in mere figures is so plainly impossible that there must be endless scope for well-paid advice on how to do it. But don’t you think that life with a statistician would be rather, shall we say, humdrum?”

Lamia was silent. She felt reluctant to discuss the surprising depth of emotional possibility which she discovered below Edward’s numerical veneer.

“Its not the figures themselves,” she said finally, “its what you do with them that matters.”

我想，這段話已十分有力地為統計的工作作了註解。

最後，我選一段臺灣作家的作品，讓大家欣賞一下。

離開了卓教授的宅子，站在陽明山的雨夜裡，我找不出下山的方式，沿路上不見任何計程車，走了許久，也未見公車站牌，直到一輛轎車在面前停下，駕駛是個三十來歲的男人。

這個月的小雨，好像從沒真正停過。生平首度搭上了便車，只因為我看見後座的一個華納卡通金絲雀玩偶，有那樣一只玩偶的男人，該有著一顆溫柔的心吧？打開車門時我想到了近日轟動的社會新聞，割腿之狼，割喉之狼，中山之狼，計程車之狼，這簡直成了一個步步殺機的城市，而我是一隻練舞的小羔羊，但天雨不斷，我上了車。

男人問清了我的去向，提議送我到敦化南路底，他就啟動了車子，又嘎然停車，在我的緊張戒備中，男人解開安全帶探身到腿下，取出一個紙袋。

紙袋中是兩杯飲料，男人解釋說：「本來想喝咖啡，又想喝奶昔，沒辦法決定，就兩杯都買了。正好妳來挑一杯吧。」

兩種都是我不常喝的東西，因為怕甜。為著禮貌，我挑了咖啡。男人果然是溫柔的，只是多語得驚人，開著車，他就有始無終地說起話來，現在他將自己細說從頭。

男人大學時從生物系轉唸了國貿系，畢業之後，順利地考取了公務員，從此在一棟四季吹送冷氣的宮殿裡上班，屬於行政院裡某個掌管統計的單位，說到此處，他自動插播說今天上山是朋友聚會，然後繼續原話題，男人在年少時夢想著的非洲人猿、紅毛猩猩和剛果金剛，抽象化成了數字、數字、數字，說到這裡他就笑了，「有時候看著看著，覺得阿拉伯數字2還真像抱著幼仔的獬獬呢。」他說。

男人負責統計，統計各種物價指數、失業率、進出口成長率、外銷訂單統計、工業生產指數、民間投資成長率……再加上物價波動預估、

國際局勢展望、重大政策效應研究，總體的目標是經濟成長率統計，然後將所有數據昇華成景氣燈號，偶爾也換個角度，算出某種叫做國民痛苦指數的東西。

爬在數字間，是純理性非感性的工作，男人這麼解釋說，但是只要事關統計，一定牽涉條件前提設定，那才是數字遊戲奧妙之所在。數據來自民間，前提來自層峰，而層峰感性得奧妙之至，所以男人的工作漸漸地偏離數學，傾向美學，他與同僚們按照指示處理數字，才在上個星期，做出了本季景氣黃藍燈的報告。

「但是有時候不管我看什麼都像泡沫，越看越像，妳看這杯奶昔根本就是泡沫嘛，妳那杯也像。」

這終結了我應酬的興致，男人於是又自轉了話題，等待紅燈時他掏出皮夾，抽出其中一張護貝照片，半帶著靦腆說：「這張照片一直隨身帶著，沒事就看一眼，妳要不要看？」

我看了，藍得要滴出顏料的天，貧瘠的黃地，一條泥路蜿蜒向前隱沒在地平線，路的遠方有些純白色但似乎見不到窗子的建築。攝影技巧並不算好，基本上這是一幅意境薄弱的作品。

「土耳其。」男人接回照片，放回皮夾前深情地又瞥了它一眼。

「你拍的嗎？」

「對啊，是我拍的。」男人說：「上次跟團去旅行，拍了不少照片，回來沖洗出來以後，我就注意到這一張，是在一個銀器市場外面不遠拍的，那時候人站在那裡沒什麼感覺，後來我一直看著照片，一直看，開始就想了，這條路再往前走下去，會到哪裡？那裡的人都在做什麼？有人就說我無聊了，只是很奇怪，這張照片我看得越久，就越像中邪一樣，很想走進那一條路，走下去，就這麼一趟，只要讓我走到盡頭，讓我知道盡頭有些什麼，我就滿足了，有時連作夢都夢到我走在那條泥路上，就算無聊吧，但是我真的就是這麼想。」

「那你就去啊。」

「也許喔。」男人喝了口泡沫狀的奶昔，說：「也許走完就發現，沒什麼意思也說不定。」

男人至此終於沈默了下來。

這篇文章來自朱少麟作的暢銷新作：「燕子」pp.199-201。非常有趣地描寫一個統計人。雖然，我幾乎不認識這樣的統計人。但我想，我要樂觀得多。在座有不少人去過土耳其，土耳其其實很好玩，統計也不是沒有意思，而我們也不會沉默。

七、品味、執著和基本能力(數學傳播，二十七卷四期)

趙民德

(本文是民國九十一年四月十三日作者在高雄大學給資優生的演講)

早先，題目是「品味、執著和關心」。後來覺得這年頭關心的人真是少了，但基本能力卻日益重要，所以修改了題目少許。意思是，這年頭，都要自己靠自己。先把自己弄好些吧，至少不要把自己弄成社會的負擔。把自己弄好，先從自己的基本能力做起。

我隨便說，你們隨便聽。架構十分明白：那就是說，今天的講話沒有架構。只是東一點西一點的零碎。

三原色原理

你們知道三原色原理嗎？多半的資優生好像應該知道。

所有的顏色，都可以從三種基本色配出來。這三個顏色稱為光的三原色(RGB)，就是紅(red)，綠(green)，藍(blue)。所有的彩色電視機、螢幕都具備產生這三種基本光線的發光裝置。因為這三種光線的混合幾乎可以表示出所有的顏色，因此電腦裡頭就用 RGB 三個數值的大小來標示顏色，每個顏色用 8 bit 來記錄，可以有 0-255，共 256 種亮度的變化，三種乘起來就有一千六百多萬種變化，這也是我們常聽到的 24 bit 全彩。

這是從網路上抄來的。我今天提起三原色，和電腦繪圖全無關係。我是說「基本能力三原色」。那就是：本國語文要夠好，至少要能輕鬆使用一門有用的外國語，還有，高中的數學要好，至少不怕。有了這樣的基本配備，就差不多可以靠這三招兩式就闖江湖了。

我說要精通母語，這裡完全沒有本土化的意思，今天我全然沒有意識型態。據說人在三歲以前，就把基本的語文能力建立了。如果把人腦用電腦來比，母語就是你的機械語(machine language)。電腦是靠好的硬體，同時也靠好的軟體。你們現在是資優生，那表示至少硬體方面不輸別人。你最早學會的言語就是你的母語：你用它來說，用它來寫。更重要的，你用它來想。——簡單的、複雜的、理智的、感性的，都用母語在運作。如果你不能精通自己拿來思想的最基本的工具，你老早就輸了。——完全是輸在起跑線上，保證外面所有的才藝班都救不了你。

所幸，我們的中文是一種豐富的語言。中文所代表的文化也是深刻廣泓的文化。中文的複雜度和豐富的詞藻和語意，讓你能夠深刻的思想，也讓你能夠表達深刻複雜的思想。而你在那裡所吸收的養分，也能讓你不會因此輸給別的語系，不會因為你身為中國人、需要用中文做母語而輸在起跑線上。會輸，是因為你不曾用心去做中國人。

如果你生來是用英文、法文或者日文思考的當然另當別論，但這一局棋如果不在你三歲前就已定局，至少在你小學畢業時就已成定局。這是已深深嵌入了你最深層的、最靈活，最能永久記憶的 CPU。人腦到現在還沒有「洗掉重灌」這一招，也許有一天你可以植入晶片，讓你忽然多了一個 giga 的有用資

訊。但是，到目前為止，你還是得靠三歲或者小學以前就植入的記憶自己努力。

用 850 個英文單字，理論上可以足以表現任何英文能表現的事情和觀念。但是你大概必需是蠻厲害的英文專家。這是由 C. K. Ogden 先生研發出來的 Basic English。

用這 850 個單字來寫書，理論上是可行的。——但「簡易聖經」還是用了 1000 個英文字，並且（想一想也知道）這是由英文好得冒泡的人執筆的。——用不夠豐富的工具來做高級工作，需要極高明的功力，可不是想像那麼容易。武俠小說裡只有功力絕高的高手才能「飛花擲葉，俱能傷人」。一般江湖混混還是用 AK47, M16 或者烏茲比較有效。別聽政客的話——不管是哪一黨的。你對 machine language 的選擇，絕對要「西瓜靠大邊」。

下一步是第二語文。這是我需要承認的無奈：誰要自己的國家不夠強呢？你需要透過另一種工具把頭伸到外面去透氣和吸取更多的養分。第二語言是視窗，透過它你和所有的網站連線。別人的本事你要正視——除非你什麼事都願意慢上一拍半拍的，等別人替你翻譯。那麼你也許不是輸在起跑線上，你會輸在等候或是和「聯外」有關的每一步。起碼，對你的專業、第二語文要精通到「能輕鬆使用」的程度。

你們是數學資優班的學生，高中數學我自然不擔心：你們懂得的邏輯思維方式，比如說什麼叫證明，什麼叫矛盾，應該夠了。這三種基本能力，也許不能讓你功成名就，但是，好好活下去應該不是大問題了。

人其實不只為了活下去

人其實不只為了活下去，人生裡要求的是要有所感動。感動所觸到的是你的靈魂，而這很難能由世俗的、靠普通程度的努力而得到。它是信念和執著。對，執著。

女兒前一陣在看日劇「將太的壽司」，被她盤據著電視和 RC，也因此看了一點。才忽然覺得其實在任何一個行業裡，都有瀟灑珍貴的一些情懷，後來在網路上，又讀到一個年輕讀者的留言：

對「將太的壽司」中的一些感動...這是摘自「將太的壽司」，東京都大賽結束後，將太的對手奧萬倉對自己表現的檢討：

「我好丟臉.....」

「不是為了輸了比賽而丟臉，問題不在結束...，我是對自己生氣...!」

「自以為是刀法名人而目空一切.....甚至忘了去看到更多自己該做的事。第一次看見關口是在比賽的開會式上，那樣小小的身軀，不自在的神情，樣子就像還沒有長大的孩子。沒想到在第一回的刀法競賽，他以努力與我打成平手，甚至現在還與清水雙雙成為決勝者.....」

「我根本沒有時間驕傲自大，應該要盡量超越自己，朝更高的目標努力才對.....」

「幸好.....比賽並不代表終點。當我在無止境的修業道路上再遇上關口，等那個時候我向自己發誓，我要超過關口！」

這原先是一年多前我的一篇週記內摘錄的一段話，當時是想勉勵自己不要因外在的掌聲而自我膨脹.....

如今回頭來看看這一段話，真覺得「將太的壽司」是部能鼓勵人的好書。

無論是奧萬倉的話，高田在聽了岩崎老師的往事後才發覺自己的努力微不足道，或將太的父親雖有個高超技術的兒子。但仍懷著不斷力求進步的壽司師父魂

魄，以及將配料也能有絕佳風味的道理傳達給北岡先生的飛男……

這些故事是自國小三年級以來我的一些感動之處，而這些已形成支持我面對種種比賽、考試的精神力量。暑假起我就是高三學生了，將來的一年是我全力衝刺的重要時期，希望我能像將太一樣，穩紮穩打地將努力化為大考後的喜悅！

(Origin: 蕃薯藤網托 From: 163.32.247.1291)

我雖不愛壽司，卻也因此而知道「鳳壽司」呢！

閒書裡得到的信息最好

我們在閒書裡得到的信息，有時比正書裡得到的多。也許是當我們在休閒的時候，心中最沒有防衛，因此作者想給我們的信念，比較容易輕易地穿了進來。也許是作者苦心經營那樣的場景，也許是因為那些作者畢竟是說故事的能手，總之，我們就一半清醒一半迷糊地學會了。

發現日本美食電視節目特別喜歡一種題材：

(頑固老闆開的好吃餐廳)

「我是開麵店，不是開酒店，想看笑臉到別的地方去！」

某一集裡，某位老闆在接受訪問時臉若寒霜地說了這麼一句話，氣溫瞬時低得連攝影機都因而微微發起抖來。

「被罵也沒關係，我好想吃一次這樣的美味……」

瞠目結舌之餘，我每每由衷地發出這樣的感嘆；我想，這種對於料理的極致追求態度，才是頑固老闆們在日本永遠備受愛戴的真正原因。

(<http://yilan.url.com.tw/all/all-010130.htm>)

其實，這樣的吃客也是蠻有貢獻的。

天涯何處無芳草

走到忠孝東路七段底，你就看到老張的炭烤燒餅，他賣的是夾肉胡椒餅、正宗的江北炭缸燒餅和蟹殼黃等等。東西種類不多，但是他一開爐總是有人排隊。早先不敢說，近十年來都是這樣。

老張的燒餅有多好呢？我不是葉怡蘭，寫不出來，但是老張是認真的人，認真而十分考究。「這一爐火候不好，不賣！」。我是欣賞這樣的人，比欣賞他的燒餅還多些呢！

下面的文章是葉怡蘭的。她的網站頗有可觀——我靠它找好吃餐館（雖然有的蠻貴）。其實我喜歡的是她能把吃東西這一件事，寫得深情款款。

(北宜公路 16.2 公里的林老闆)

前往群山環抱裡的這家小小獨棟餐廳那日，秋日裡微寒的毛毛靈雨已經接連地下了一整日了，然而，店裡依舊暖烘烘地高朋滿座，室內佈置十分溫馨，榻榻米、實木厚桌、扶疏的花草盆栽與各種古樸擺設收藏相映成趣，流露一種非常雅緻隨性的居家味道。

菜肴部份最令人難忘的，是那極具水墨畫意的排盤方式：例如一方大陶盤裡，以一段梅枝、幾片新葉、些許留白圍繞著一碗蘋果山藥沙拉，或是一鉢連花蓮藕排骨燉湯上頭，一整朵蓮花盈盈盛開著；每一次上菜，都引來一回驚嘆。

而菜色本身，則也同樣充滿著這樣率性揮灑但創作概念十足的氣氛：清新的苜蓿芽山藥卷、綿細粉嫩的手工豆腐佐山葵、幾滴香橙酒醃過的生魚片、香鹹微辣的泡菜粉蒸排骨、濃郁有味的蘿蔔納豆蒸魚……，都是非常獨特、帶有一點點靈光乍現味道的奇妙組

合。

——讓我聯想到一些著迷於庖廚的設計師、畫家、作家朋友做的菜，不管是主是客，不管是烹調是品嚐，都彷彿經歷著一段充滿創新與試探的冒險。

遂而滋味上總是格外有著一種快樂而驕傲的自信自在，與氣氛裡不時流動著的熟稔相契兩相呼應，相對成為這一類餐館之所以永遠讓人流連再三的最主要原因。

第一次別人叫我去的時候，先還以為只是一家土雞城。現在卻成為太太較愛去的店之一（其實她不愛生魚片）。我們告訴自己不要去太多，免得嘴刁了以後平常的日子難過。但是那一晚，我們帶著羅小華去——他是那種先走完一遍小吃街再決定選哪家店的人——我們一路驚嘆不已，最後，林老闆都出來了，還拿出紀念冊要我們題字（這時母語要好就有用了），林老闆雅得要命——每一盤上來的菜，都是「極具水墨畫意」，更別提大堂裡的古雅擺設，清新忘俗的畫作和楹聯了。如果你只寫了「好吃、好吃，真好吃……」，雖然真切，恐怕林老闆會覺得我們這些理工科教授其實沒有學問呢。

那一晚，我們吃得太飽，我只題得出「此間有真味，欲辨已忘言」這兩句。這裡面 90% 是陶潛的。回來後覺得不夠意思，努力完成了一首五絕：

四圍皆綠樹，欄外是青山，
此中有真味，欲辨已忘言。

這樣給林老闆看，就應該差不多了。

老虎、老虎、老虎！

我其實愛看戰爭小說。一本一看再看的，是黃文範先生所譯的「中途島之戰：難以置信的勝利」。這是我信得過的譯者，因為他一絲不苟。

五十六年秋天，我在台南市買到了華特勞德所著的「中途島之戰：難以置信的勝利」，讀過以後，深深為它生動、細密的筆調所吸引，便有心加以翻譯；可是卻遭遇了最大的困難——把書中英譯的的日本人名、地點、和艦名再譯回漢字。我翻遍了字典、參考書、和類書，請教過老師和朋友，謝謝內兄傅琦經過東京時，還特別為我購得牧島貞一的「中途島海戰」，但是仍有一部份譯名不能確定。

譯史——尤其是譯日本人的地名，「信」最重要，既然無法完全譯出，幾次都擬擱筆，但本書引人入勝，又實在割捨不下，只有盡其在我，把一切疑難譯名做成卡片。然後，就像書中的美國特遣艦隊般「等待！只有等待！」

感謝我在台南亞洲航空公司時的同事——目前遠在夏威夷的張凱玲，她是位虔誠的基督徒，希望我能「得救」，不只一次要我在公餘去教堂聽道。五十七年七月上旬，她又勸我去聽一位日本傳教士在台南市的佈道，她提到這位長老會牧師的姓名，使我霍然振奮，真想想不到居然有這種難得的機會，他就是近代史上赫赫有名的淵田美津雄。

淵田是珍珠港空襲的總領隊。後來因為受了感動，做了牧師，這是閒話，表過不提（見 http://www.mediaconnection.org/press/020204_7.htm#7）。但是為了幾個日文譯名這樣的用心處理，和林老闆特意去南投山上找一對老夫婦所做的梅子醋來配菜，還有南港的老張總是堅持用最好的里肌肉做他的胡椒餅，精神上都是可愛的執著。

執著是成功的要素。執著，是世俗化的第一步。有趣的是：這並不一定是壞事。

中神通的問題

有一天和太太去見陳省身先生，陳先生是數學大師（真的大師，不是記者隨便封的大師），學問是不必說了，其他的做人、做事、做長輩……也是極有風範。傳承德好些年前有一次替我帶禮物給他，回來之後感嘆說：這才是「中神通」的樣子。

在研究院上班，東邪西毒南帝北丐之輩是見過一些。但是俠骨英風的王重陽，卻極為少見。那是一種獨領風騷的大工大將，再加上洵洵然的儒者之風。其實也難怪，有一次聽他提起：他是閒來讀資治通鑑的。

資治通鑑是司馬光寫來教皇帝怎麼當好皇帝的書——如果誰堅持用河洛文，他絕然看不懂這樣的書，因此也絕然學不懂裡面的道理——閒話休提，陳先生是愛吃「鼎泰豐」的。於是話題便轉到中國菜。

「外國人曾問過我哪一樣中國菜是特別值得大書特書的。」他說：「你們怎麼看？」

這時就要靠我的武俠小說讀得多了。我說：「二十四橋明月夜怎樣？」

（二十四橋明月夜）

當晚黃蓉果然炒了一碗白菜、蒸了一碟豆腐給洪七公吃。白菜只揀菜心，用雞油加鴨掌末生炒，也還罷了，那豆腐卻是非同小可。

先把一只火腿剖開，挖了廿四個圓孔，將豆腐削成二十四個小球分別放入孔內，紮住火腿再蒸，等到蒸熟，火腿的鮮味已全到了豆腐之中，火腿卻棄去不食。洪七公一嘗，自然大為傾倒。

這味蒸豆腐也有個唐詩的名目，叫作「二十四橋明月夜」。要不是黃蓉有家傳「蘭花拂穴手」的功夫，十指靈巧輕柔，運勁若有若無，那嫩豆腐觸手即爛，如何能將之削成廿四個小圓球？這功夫的精細艱難，實不亞於米粒刻字、雕核為舟，但如切為方塊，易是易了，世上又怎有方塊形的明月？

（金庸：《大漠英雄傳》）

用這樣的精品去唬弄只知道「李鴻章雜碎」的洋人，可說是殺雞用上了牛刀。這樣的豆腐，我自己都從來沒有嚐過呢！

回到司馬光，我還一直以為他是唐宋八大家之一呢，幸得一位同學提出疑問。——數學資優的同學，果然國文也不錯。

讀司馬光可以造就大學者的風範。其實讀凌解放（二月河）也不見得就沒有用。過去二十年羅小華每次回台，除了看父母就是來我辦公室拿康熙大帝、雍正王朝…有用沒有？我可是看著他從助教授一直幹到系主任呢！

落花時節的溫柔

當然，這裡面還是有一點品味的不同。現代的理工教授國文是差得多了。品味是要在行有餘力的時候才能有的。黃易在他的小說裡常說「發財才能立品」（這是杜伏威向寇仲解釋為什麼他的江淮軍形同流寇的道理）。品味兩個字很難說出來。高陽的小說裡也常說「三代為官，方知穿衣吃飯」，並時常笑內務

府（皇帝的總務）官員：「樹小、房新、畫不古，此人必是內務府」。總之，品味是和暴發戶行為是相反的就是啦！

講到品味，我喜歡比較兩首詩。例如白居易在琵琶行裡有下面的名句：

同是天涯淪落人，相逢何必曾相識...

這是描寫一個謫官的詩人在江邊船上遇到一位已嫁作商人婦的女士並聽了她的琵琶之後的感嘆。這當然是千古名句，但是白詩「老嫗都解」，因此意思十分明白，明白到把感覺似乎一下子抽光。以至於到了最後：「座中泣下誰最多，江州司馬青衫濕」，把餘意一再壓榨。詩還是好詩，但味道因為感情有些浮面，就比下一首杜甫的「江南逢李龜年」弱了。杜甫的詩只有幾句：

岐王宅裡尋常見，崔九堂前幾度聞。
正是江南好風景，落花時節又逢君。

沒有說一個愁苦之詞，落花時節的相逢，對比著以前的交往。這樣的風流蘊藉，情在言外，真不愧是詩聖的作品呵。

五次方程式之解法

說過了兩首詩的味道，讓我們再回到數學家的故事，這一次我們說華羅庚。他在未被發現之前，在鄉下的一個學校做「庶務」的工作。故事由蘇家駒先生寫的「代數五次方程式之解法」（1930《學藝》v.7, no. 10. 上海）談起。

國中生都知道一元二次方程式的公式解。我們以前在高中時所學的「范氏大代數」（我們管它叫「繁死大代數」），還有一元三次方程式的代數解，已經足夠麻煩了。四次方程式也有公式解，但我就沒有在課堂上學過了。五次呢？

一般的五次方程式是沒有公式解的（不是說無解，它當然有五個根，但我們沒有已知的公式可以硬代的）。在1930年代，知道此事的固然有，不知道的也不乏人，知道了還不信的人，自然也還是有。

這位蘇先生就發表了這麼一篇論文。華羅庚那時還沒有出名，讀了這篇文章，也做了一篇，且看他的論文摘要：

五次方程式經 Abel, Galois 之證明後，一般算學者均認為不可以代數解矣。而「學藝」七卷十號載有蘇君之「代數的五次方程式之解法」，羅對此欣喜異常，意為果能成立，則於算學史上亦可佔一席之地也。

雖自思若不將 Abel 言論駁倒，終不能完全此種理論。故羅沉思於 Abel 之論中，凡一個月。見其條例精嚴，無懈可擊，後經本社編輯員暗示，遂從事於蘇君解法確否之工作，于六月中遂得其不能成立之理由。羅安敢自秘，特公之於世，尚祈示正焉。

請留意他的態度：先花一個月了解基本的道理。並且，直接找最重要的原始論文去研讀。直到他認為這兩人的理論不錯，才努力去找蘇文的謬誤。——這是一個非常有邏輯性的學術歷程。並不是隨便就坐下來解題，解完了也不知道對不對就試著去發表。

蘇家駒先生是另一種典型。也想去解難題，但只想憑著眼前手上的功夫。

他的錯處必然也是不那麼明顯的，但他愈錯得不明顯，便愈顯得華羅庚的文章如何的深刻——說白了，以小有身份的蘇家駒的錯，便足以提供華羅庚以出人頭地的機會。

果然是一個機會。這文章被熊慶來和楊武之兩位北大教授看到，同時都起了愛才之念。那個時代，大教授說話還是有人肯聽的，他們排除很多技術上的困難（華只有初中畢業），把他弄到北大和陳省身同時做助教。華當然沒讓這兩位教授丟面子，後來變成世界級的數學家。

70 年末期，華羅庚和他的學生王元教授訪美。我因為認得王元的弟弟和妹妹，就設法安排了讓 Bell Lab 接待了他們，而我負責招待。那時手邊正恰有兩冊華著的「數學分析」因此得到他的簽名。這兩冊書一直放在我的架上——直到今年。

今年傅承德指導一位資優生，他認為是一個天才。我們聊著怎麼給他最好的機會。

「給他好書讀，別急著做研究」是我的想法，這個時候學什麼都記得住，花時間做二三流的問題以得一個科展獎，是有點浪費了。

我本著「寶劍贈烈士，紅粉贈佳人」的心情，把兩本有華羅庚簽名的書送給這年青人。也許他會去學數學，也許不會，但那並不打緊。讓他讀一些大師級的數學家的書，比做考古題有意義多了。

為什麼呢，這是品味的問題，年輕時千萬別把品味搞差，將來再也救不回來。品味不只在歷史書和歷史故事出現比較、不只在對的論文和錯的論文間比較、在好餐館和差勁餐館之間比較、在好詩和更好的詩裡比較。你知道嗎？我的孩子說伍百成名之後的音樂反而差了，陳昇也是。——維持品味其實並不容易。雖然，我在前面說過，丟掉執著有時也並不是壞事。

我也看席娟

再舉一個比較，你知道席娟嗎？下面的場景是她的小說中安排的男女相遇：

突然，黃耘春的腳在桌上踢了她一下。

「做什麼？」

「靠門那一桌的那個男人一直在看你，我保證打從你進來後，他一直沒把眼光移開。」黃耘春小聲的對她使眼色。

這是身為女人的驕傲。也許有些女子在表面做懊惱無聊狀的同時，內心會被虛榮心吹得飄飄欲仙！

可是，那不是她。原穎人自知長得有些媚，看來很好勾搭的樣子，也就因為如此，打十四歲開始就吸引了一些不入流的人種來搭訕。

在那些人的眼中，她容易上手，可以玩一玩、逗一逗，而絕不會有人認真的待她。有了這項認知，她更應該目不斜視、端莊自重，否則只消她斜眼一瞄，保證人家立即當她拋媚眼，到時可真是百口莫辯了！

真不知是什麼道理！她全身包得連蚊子都會給悶死般的緊密，怎麼還會有不肖變態男盯上她呢？

是的，她想戀愛，但絕不是在隨便的情況下、隨便的天雷勾動地

火！全世界有二十五億男人，全台灣有一千萬男人，她不必太著急，更無需因一個無聊的盯視而在心中想得亂七八糟！

她是作家，要端莊、要有氣質、要特別，尤其挑對象更要找一如自己筆下寫出的情景，才不會讓讀者們幻滅！這條款是……唔！作家的面子問題。

「喂！看一眼嘛！長得挺正點的。」黃耘春又踢了她一腳。原穎人卻故意將右側的秀髮垂落到胸前成帘幕，阻隔一切可能的注視。會對一個陌生女人行注目禮的男人會是什麼好東西？不是色鬼就是登徒子！

「穎人！我這間小店的客人都挺有水準的，那俊男也許是某一棟辦公大樓的白領上班族，全身上下都是雅痞的味道。相信我，這麼好看的男人很少見的，交來當男友會很風光。」

席娟：《女作家的愛情冒險》

席娟的這一本書其實蠻俏皮的。比她其它寫來騙國中女生的小說要好。那一類書，有很大一部分是因為市場需要而搞糟的。不能怪席娟，只能怪這個市場。但是，你不妨讀一讀侯文詠怎樣來處理類似的場景。

認識關欣時，他們都還是學生。

當時很流行醫療服務性質的社團。學校教授找來一些研究經費，動員醫學院學生到偏遠地區做醫療服務，同時做一些公衛方面的學術調查。關欣在醫學院低蘇怡華二屆。他參加了訓練營才認識她。

那次訓練營有一堂令人打瞌睡的課，他無聊地在筆記本上塗鴉：

這次我離開妳，便不想再見妳了。

妳笑了笑，我擺一擺手

一條寂寞的路便展向兩頭了。

關欣正好坐在隔壁，頑皮地加進來接龍：

念此際，妳已回到濱河的家居。

想妳正在整理長髮或是濕了的外衣。

他們玩得很開心，發現兩個人幾乎可以背誦大部分鄭愁予的新詩。

說不清楚那時候為什麼很多照片的背景都是海。有一回他們騎著摩托車夜遊，騎到石門海邊看漁火時已經半夜了。他們並肩坐在海邊聽濤聲。或許是風吹起她的頭髮撩撥蘇怡華的緣故，他側過臉吻她。

那算是他的初吻。吻完以後他自己都不知道為什麼會這樣。關欣站起身來，在沙灘上走啊走地。蘇怡華心神不寧地跟在後頭。走了不曉得多久，關欣才回過頭，像唸詩一樣，輕描淡寫地說：「如果你喜歡海喔，就不應該試圖靠她太近。」《白色巨塔》

比一比，你知道味道是如何的不同嗎？我竟不知鄭愁予的詩是這樣溫柔的！

井中之月

夠了，品味是什麼呢？對我來說是年紀越大，口味越淡。回首千山路，昔

日的追求和執著，有時也和雲煙相仿。到我這個年紀，可以說和尚的故事了。

有一個和尚過河的故事是這樣的：話說有 A, B 兩和尚，正擬涉水過河時來一美女，說怕濕了衣裙，煩請師父捎過河去（有沒有搞錯？）。A 同意做，因此無話。

美女與 A, B 都平安過河。A 把美女放下，美女道謝離去。

A 和 B 續走了 20 里，一路無言。B 忽然對 A 說，師兄，當時你實在不該捎這女檀越的。

A 說，我過河後，已將她放下，為何你至今還不曾將她放下？

（這故事我已將它放下，因此不記得出自何處。）

人世間事，不外乎拿起、放下。只是不知道該拂去塵埃還是不管塵埃罷了。但是全心投入，有時反而心無掛礙，旁觀者卻反而容易動心呢！過河只是過河，打電玩便只是打電玩。下面是方鏞欽的故事。

（全心投入）

我彷彿是網際網路虛擬的「網路運動場」中，創造出來的第一個「最有價值的玩家」，但是這個頭銜的由來，則是一場令我終身難忘的比賽。那就是九八年的職業電玩 PGL (Professional Games League) 全美比賽，

雖然我和隊友取得了團體賽冠軍，另一個更艱鉅的挑戰正等著我，我在個人賽決賽中的第一回合，輸給了強敵「不朽者」，我必須連贏兩場，才能奪標。然而，難纏的「不朽者」僅需要贏一場就登上全美矚目的冠軍寶座。

我和對手認識五年了，在這場大賽中，我們第一回合交手時，可能是太熟悉對方，我沒有集中注意力，竟然輸了第一回合。我大方地告訴他："Great Game!" 別人為我緊張，擔心我會因此一蹶不振，信心全失；可是我並不緊張，也不覺得挫折，還是信心十足，只是對自己感到不滿。因為我知道自己沒有盡力，我一直告訴自己，要再集中精神、再專心，Just do it!

又遇見這號人物——「不朽者」，我的電玩迷為我冒汗，而我自己當時腦子什麼也不想，就是閉目養神，讓自己冷靜，再冷靜，來自心裡的聲音對我說：「盡力就能成就一場最美好的比賽。」在接下來的時間裡，我的眼、我的心、我的手，像是與螢幕的一切融合為一，我以十比三贏了這一局。

這樣戲劇化的結果，眾人響起震耳的驚呼聲。這些激動的情緒，當時，我真是聽不見，也看不到。我只是和「不朽者」打平了而已。

勝敗之爭才正開始。在接續的爭霸戰中，就和往日的關鍵時刻一樣，我有一種「通電」的感覺，通往一種神奇的能力，我真的能感應到對手下一個動作，下一個戰略。我先發制人、迎頭痛擊，出奇制勝。二十分鐘之後，勝負了然於心——四十三比零。

我的支持者驚呼不已，怎會不興奮，我覺得自己戰勝的不是任何人，而是我自己。

（取自《玩性堅強》，圓神出版）

這正是黃易小說裡的「井中月」的境界（你應該抽空去他讀的小說：大唐雙龍傳）——平心靜氣的一往無前，因為專注，所以心同井中之月，關照一切。無他，無我，和遊戲成為一體。

心中但存一念

天下事難嗎？容易嗎？端看一心。

(禪＝吃飯，睡覺)

有緣問珠海：和尚修道，有秘法否？

有。

如何秘法？

肚子餓時吃飯，身體睏時睡覺。

與常人吃飯睡覺何異？

不同。

敢問？

彼吃飯時百般挑剔，嫌肥揀瘦，不肯吃飽；睡時胡思亂想，千般計較。
(這是從星雲的書上看來的，我只是將它譯為文言。)

別問為什麼，當你沒有懷疑的時候。科學的精神，是設法從複雜的現象裡，剔去噪音，只留下本質；而世人卻喜歡凡事都泛政治化。

傷心時不要跳舞

(蜈蚣是跳舞的高手)

「這是一個非常嚴肅、非常哀傷的故事。」

「說吧。」

「從前有一隻蜈蚣，可以用牠那一百隻腳跳出非常美妙的舞蹈。每次牠跳舞，森林中所有的動物都會跑來觀賞。大家對牠那美妙的舞姿都印象深刻。可是有一隻動物並不喜歡看蜈蚣跳舞，那就是烏龜。」

「牠大概是嫉妒吧。」

「烏龜心想，我要怎樣才能阻止蜈蚣跳舞呢？牠不能說明他不喜歡看蜈蚣跳舞，也不能說自己跳得比較好，因為那是不可能的。因此牠想了一個很狠毒的計畫。」

「什麼計畫？」

「牠坐下來寫了一封信給蜈蚣，說：『喔，偉大的蜈蚣呀，我對你精湛的舞藝真是佩服極了。我很想知道你是怎麼跳的。你是不是先舉起你的第二十八號左腳再舉起你的第三十號右腳？還是先舉起你的第十七號右腳，再舉起你的第四十四號右腳？我熱切的期待你的回信。崇拜你的烏龜敬上。』」

「真是鬼話！」

「蜈蚣讀了信以後，馬上開始思索自己是怎麼跳的。牠到底先舉起哪一隻腳？然後又舉起哪一隻腳？你猜後來發生了什麼事？」

「蜈蚣從此不再跳舞了？」

「正是如此。這就是理性的思考扼殺想像力的例子。」

「這真是一個悲哀的故事。」

(《蘇菲的故事》)

這故事的教訓是，不要跟著別人的節拍跳舞，對蜈蚣是如此，對你也是如此，擔心別人的想法是會老得快的。

等待最苦

有一個緊張型的男子，樓上住著一位女郎。每晚回家，女郎先脫下高跟鞋，「拍」一聲，一隻高跟鞋落在地板上，再過一兩秒，「拍」，另一隻高跟鞋也落在地板上。

男子甚為所苦，有一次終於找到機會，婉轉表明了「請晚回時別用高跟鞋丟地板」。

女郎是受過好教育的知識份子，「沒問題，」她說。

第二天夜裡，女郎又晚歸如常。腳一抬，「拍」，高跟鞋一隻落地。

「噢，」她想：「糟了，吵到樓下。」於是小心翼翼地輕輕脫下第二隻鞋。你猜怎地？——還能有第二個結果嗎？樓下住的神經質的男子，整夜不寐——他一直在等不知道第二隻鞋什麼時候會落下來！

(這是蘇競存三十年前告訴我的故事)

我繞了一個大圈子，一路說些閒話，有些和數學有關，有些一點關係也沒有。當然啦，我既不想講做人的道理，也盡量少講做數學的道理；只是想說些故事讓你們想一想。意思是說：好像講了不少，但其實沒告訴你們甚麼。

你聽過會跳舞的小狗的故事嗎？這是最後一個故事了，我在網路上看到。
(會跳舞的小狗)

有一天有一個人帶著一條狗到唱片公司，他說他是這條狗的經紀人，並說他這條狗會唱歌跳舞。

老闆不相信，就叫小狗表演一次。

當音樂響起，小狗跟著音樂載歌載舞，老闆目瞪口呆的看著小狗，一邊想著這一次撿到搖錢樹了，就趕快拿出合同希望與狗簽約。

沒想到忽然一條大狗衝進來，把小狗銜走了，老闆問：「怎麼回事？」

經紀人無奈的表示：「唉！那是他媽媽，他媽媽希望他兒子成為一個醫生，演藝圈太複雜了。」

其實世事都是複雜的，學術圈子裡面並不缺狗皮倒灶的事和人。我們只是用一點心力，希望能造就幾個將來的數學家，安知你們的媽媽不希望你們去做醫生呢！

(這是稍早些時的講稿，我將它由 Power Point 檔，加了一些工之後，再轉成文章。文中所提的人名，都是真有其人。如有不雷同處，都是我記錯。)

——本文作者任職於中央研究院統計科學研究所——

八、萬物有常・世事多變(2001)

趙民德

中央研究院統計科學研究所

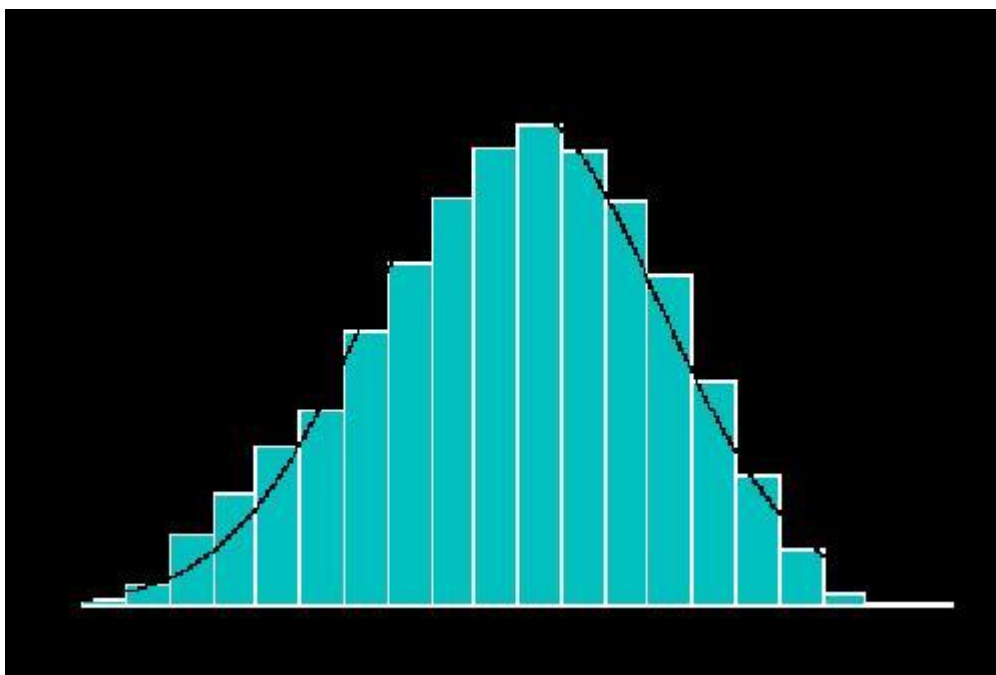
曾經有一支李壽全的歌，也許現在還在 KTV 中可以點到，大概是叫做「我的志願」。歌詞的意思說，小的時候，他的志願是做總統，做發明家，後來日漸長大，才知道總統只有一個，真正的發明家也不太多。志願在聯考時填一百多個之後，也就不成其為志願了。而當年紀更大了，才知道天下的事情都是正正常常地運作的，大多數人都是普通人，既不是總統，也不是發明家。雖不是混混沌沌地過日子，但是快樂是自己定義的；不快樂也是。因此我們可以悲傷，可以歡笑，和普通人一樣。

這個世界本來就是這樣的。不管你自己覺得你是如何的了不起或特立異行，但大家都是在同樣的環境之下培養出來的，吃同樣的米，喝同樣的水，呼吸同樣髒的空氣，讀同樣的教本，看同樣的武俠小說、漫畫和港片。因此不論是在你的身上做甚麼測量——比方說腰圍吧——那麼量測出來的數字，如果放在一大堆類似的人的同樣的測量裡，就會看不出有多少不同。因為你是一個群體裡面典型的一員。典型的意思就是說：大家都是普通人，你當然也不會特別地特出。

這當然是悲哀，也是無奈，但也不是那麼的一定就是壞事。不論甚麼，我是說不論甚麼，只要不是強迫性的（例如二中選一的結果），要是可以量測出來（例如可以到小數二位）的事物，大概就是這樣。你不論去看一個羅馬時代的軍團，或是去看一個成吉思汗的萬人隊，大概就是這樣：大部分人都是普普通通的。

普通或特殊是在和類似的人相比之後的結果：多半的人一切都普普通通。因為那就是普通的定義，你若是在前面的百分之三，那你當然不那麼普通了，你若是在後面的百分之三，那你也當然不那麼普通了，但 94% 的人既不在前面三個百分點，也不在後面的三個百分點裡。百分之九十四的意思再明白也不過，在街上隨便找一百個人，平均有九十四個是普普通通的。

這裡面有學問嗎？這樣的簡單道理，粗粗一看，是沒有甚麼學問的，但用一點點畫圖的技術，問題馬上就不一樣了。當然，也就有了前、後的百分之三以及當中的百分之九十四。圖一是民國 77 年度大專聯考裡兩萬多名考生的英文成績。圖一所畫的是所謂最簡單的長條圖。



圖一

這個圖像甚麼？這差不多是但又不全是常態分布的曲線。這當然不那麼像常態分布，但也差不到那裡去，反正它基本的道理是十分清楚：極差和極好的考生之外，都是普通考生。普通者，佔大部分也。

計算一下這兩萬多學生的平均值和標準差。圖一裡的曲線是利用這兩個值所算出來的常態分布曲線：是相差不多是罷！

我們輕輕鬆鬆地，幾乎是「芝麻開門」一般地打開了一個重要的寶庫：常態分布。

- 你在中學有沒有要證明根號二是無理數？
- 你有沒有在微積分裡學過 $\sin(x)/x$ 在 x 趨近於 0 時的極限值 1，而這裡的「正弦函數」得要用「弧度制」，而弧度制，得靠圓周率等於 3.14159... 才能定義出來？
- 你有沒有學過複利，而用到 $e=2.71828...$

如果你覺得這三件事都有一點學問的話，那麼常態分布的學問可能更大，它的標準型（意思是說最簡單的）公式是

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} \quad (1)$$

我們既用到根號二，也用到圓周率，還用到自然對數的底 e 。把任何一組數目，如果它們有一個共同的題目，如果它們由測量——不論是自然的方式（如量你的腰圍），還是人為的硬性規定（如考試的分數在 0 與 100 之間），——如果將這一組數字畫一個長條圖，那麼圖一的形狀，便很有可能出來。

有甚麼用？中國人做學問，基本上上焉者是為了經世濟民，下焉者為了自己吃飯。為好奇而做的學問就少了。中國人有興趣的是用處。但純粹自用處的觀點來找重要的結果，是很難的。重要的發明或發現，原動力來自「為甚麼」，而非「有甚麼用處」。一大堆純粹因為好奇而導出的事物也許都是沒有用處的。

但偶然也有一些極有用的，常態分布是其中之一。

一個顆刻磨成 58 面的鑽石有用嗎？它可以切割玻璃不錯，它也可以討好美女，但它的價值來自它的完美：硬度無出其右，折光率也無出其右。何況研磨的技術是人類能力的尖端所能，它的價值來自它的性質。它是自然生成的，在極難的條件下生成。

對統計學者而言，常態分布的存在就如鑽石一般：它不是我們發明的，它是自然產生的，而人類用極大的心力從自然現象裡逐步提粹精鍊，而最後得到前面的公式 (1) 這樣的簡潔形式：有如一座皇冠，鑲上了 e 及圓周率兩粒閃亮的寶石，而尚嫌不足，又配了一個用根號二做的鍊子。

但這是只是形而上式的讚美，常態分布的精要，來自它所具有的性質：是這些性質讓這個分布成為一個重要的分布，而不是任何的人為因素。例如它又叫做高斯 (Gauss) 分布，但並不是因為高斯導出它它才偉大（事實上高斯的推導，還有一些繞圈子的邏輯未能釐清），它的重要來自它本身的光輝。

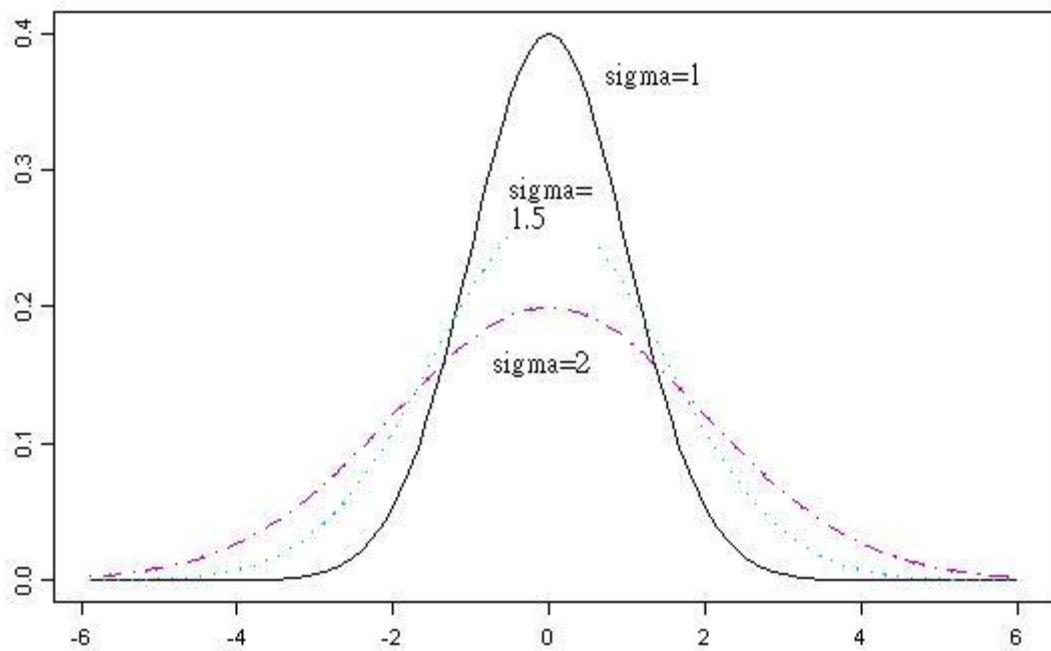
大家都知道的一條性質是：「常態分布加常態分布，仍然為常態分布。」這經常是標準統計或機率教本裡一定會講到的性質。我們現在當然不用功夫討論中央極限定理，但是這一條性質以及它的逆定理：「如果有兩種分布的和恰好是常態分布，則這兩種分布都非同時是常態分布不可。」這兩項重要的性質，左右夾擊，使得中央極限定理不止是非成立不可，並且它極限分布，也非得是常態不可。

我們先自可看到的一組一組的數據，來看是不是這樣有的分布，然後再經由一些整套的數學推導（以及近 200 年一大票學者的努力），其結論是：我們所能收集的一大堆數據裡來看，如果我們鐵口直斷地說它們是來自某一個常態分布的，那麼，我們不會太錯。

這個道理的重要就和生物學家要懂得細胞結構，物理學家要知道原子結構是相差不多的。常態分布是我們所最最了解的數據的基本結構。而我們用它來描繪人類最最不能避免的問題：誤差及變異。變異所代表的意思是同樣的物件或事情，當它一再重覆發生的時候，不論我們如何用心測量，總是不能得到同樣的結果，例如生產線上的規格，絕不可能全然一絲不少地合格。再用心挑選的儀隊成員，其身高體重都不會一樣，只是有的變異大些，有些變異小些，而變異在常態分布的時候，可以容易地用一個尺度 sigma 來表示：

$$\phi(x, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} x^2\right\} \quad (2)$$

和前面的公式 (1) 相比，這相當於公式 (2) 裡令 sigma=1 的特殊情形。我們知道 sigma 的值愈小，則變異愈小，sigma 的值愈大，則變異愈大。圖二所示的是 sigma = 0.5, 1 及 2 的三個情形。



圖二：sigma 的影響

我們當然也可以將誤差想像成變異，但我們還可以更多想一想。自從牛頓發明了第一到第三定律，我們就比較願意去想像宇宙萬事，不論是自然的，還是人文的，大概都有一個方程式可以表現出來。

這當然是一件了不起的想法，我們只要想一下力學能憑幾個公式就計算出多少道理，以由此而產生的近代科技文明，只要有一點想法的學者甚至政客，都會不禁地想：假如其他的事情也可以用公式來表示，那會有多厲害！

但世事豈能如此順心！我們才學懂高中物理的時候，多半以為世上的事情都可以用公式計算，等到年紀愈大，見得愈多，才知道人類真正能夠計算的東西，遠比他不計算的東西為少。但這並擋不進我們想去「抓住」一些現象並寄望由此能做預測的想法。通常，我們能抓住，或者至少希望能抓住的那一部分，叫做模型，而抓不住的部分，叫做誤差，而

見到的現象＝模型＋誤差

這裡，誤差若是用常態分布來表現——其實也不會非常離譜——我們就差不多概括了目前人類知道的大多數的計量模型。我們當然是將問題極度簡化，才有以上的敘述。但精神上這差不多就是目前大多數計量方法的精義。這些得過諾貝爾獎的，還沒有得但非常想得這個獎的學者，過去一兩百年甚至未來一二百年，多多少少地用這樣的基本想法來利用資料與模型，既歸納又演繹地尋找人間萬事的規則。這些工作，若沒有常態分布，免談。

人們之所以能夠這樣做，一個基本的道理是我們已經非常了解常態分布，以至於我們可以充分地透過它的良好性質，一方面安心地設立模型，另一方面由這樣的模型裡做出推論：我們對大部分天下事物的了解，並不是由於我們真的懂得其中的運行規則，倒反而是因為我們比較了解問題中的變異及誤差！

早年的農業實驗，希望能找出最佳的肥料比例以使得稻米的產量更高。這一系列的活動，背後都是基於我們對常態分布的了解而達成。至於二次大戰以後就一直不曾停止的工業品質與工程的工作，也一直在常態分布的支撐之下有

成。農業的改良讓我們吃飽，工業讓我們的衣、住和行，甚至育樂得到滿足。那麼，新藥的發展——為甚麼它們能夠既有效又無害？為甚麼我們對它的效應及害處都能夠用計量的方法加以詮註：這也得力於我們對常態分布的了解。至於人文及經濟上的活動更不必說了。幾乎可以說，沒有常態分布，而這些學問都只能停留在原始的性向描述的階段而已，那裡談得上任何解析的結果。

可是，這些模型也太沉重了一點。我們不必那樣在意國民所得，你說，那是主計處的事。我們也不必那樣在意考題的信度和效度，你又說，那是教育部的事。如果常態分布只是用來做這些官方問題，那麼我不知道也罷。

那麼你知不知道第四臺有一位老師，號稱是留美的碩士，並且用迴歸分析「這樣高級的統計方法」來解盤？

這我其實所知也是有限。但是以常態分布為基礎的統計模型，的確確在證券分析市場裡大賣特賣。其中最有名的一個模型，叫做「美國 BARRA 多因子淨值模型」。它自 1970 開始有了基本理論基礎以來，目前已使其公司達到年營業額四千萬美元以上。至少 600 家主要的投資公司（它的總資產額超過一兆美元）長期訂購這個模型。其目前所賣的 E2 模型，卻全然是傳統的統計建模（modeling building）工作。

這樣的工作有多難呢？在技術面上，並沒有想像的那樣難。如果有源源不斷而來的定期數據，如果我手上有一個像樣的主計處裡面一個技術科的人力，努力兩三年就應該有可以出售的商品了。事實上，BARRA 模型裡用了不下 60 個因子，算是一個相當複雜的多因子模型。如果只有一個因子，即所謂的單因子模型，那麼各證券的投資報酬率針對那個唯一因子（多半指某一種市場指標，例如我國的發行量加權指數）的敏感度，就是目下較高級一點的證券分析師口中所說的「貝它」（beta）值。貝它有甚麼用？它可用來測量某一證券針對某一證券組合的影響。還太難了是不是？且不必管這些技術名詞，更且不必管這些名詞的真正意義。在投資工業裡，貝它值是常年地被各家各派的高手計算以及估計著的量。「貝它服務」是可以像訂報紙一樣的按時送達的。

整個投資分析業是大量用到常態分布的行業。只是傳統的統計學者，都太帶著理學院的學者氣息，不願意真正去碰這個行業，因此，表面上它的統計意味不那樣重。但在這裡，無數個聰明精巧的心靈，極盡所能地從如海的 data 中，希望比他人更早一步找出重要的規則與信息。在另一方面，一個有效率的市場，其定義卻正是「應該無法自公眾皆可得到的 data 中找出必然賺錢的方法」的市場。這是訊號和雜訊交戰的殺場，也是一個近乎零和遊戲中永不休止的震盪與徘徊。這是一個隨機的世界，而隨機乃是常態。

人類因為有數字的概念而比其他動物高明。我們曾經以為這方面的智識是絕對的：「給我夠長的槓桿，我可以舉起地球！」但現代的人類要謙卑多了，他們知道多半的時候，力有未逮。而在他們承認了誤差及變異的存在之後，反而海闊天空地走入另一片開廣的領域。萬物有常，這是我們想抓住的部分：世事多變，這是我們用適當地常態分布來表現的部分。這個道理，好像易經裡也有。但是，將常及變放在一起，加以計算及演繹，並由此做科學的驗證，卻純然是近五六十年統計學及相關計量科學和廣用科學的發展。

我們都是普通人，這是「常」的部分。普通人也有高有短，有胖有瘦，這是變的部分。這些工作，只是儘量抓住「常」的部分，而讓「變」的部分受到控制。而這一切都希望用真實的數據和確切的計算來支持我們的論點。我們當然不只是談身高體重，我們談國民所得，失業率，國家競爭力，品質工程，大

專甄試的信度和效度；衛星信號的濾波，由報稅單裡看出異常而加以防止逃漏，環境空污……這些問題的裡面，都離不了數據，模型及誤差。而沒有我們對常態分布的了解，要想有效地處理這一類的問題，門都沒有。

常態無所不在。因為這就是「常」的意思。月有陰晴圓缺，但因為我們真正了解到沒有嫦娥玉兔和吳剛，我們反而不會失落。分子與分子的碰撞該如何計算？因為分子實在太多，我們實在無法用撞球臺上的物理學來描繪分子碰撞的軌跡。因為我們知道那樣做不行，所以我們反過來認定我們無法全然知道碰撞後的路徑，用隨機的角度來想。這似乎是破釜沉舟的法子。但這也是愛因斯坦用以算出布朗運動公式的法子。

人生因為有兩個因素而變得瑰麗：可以努力改善但又不能預測。人生又因為兩個因素而變得無奈：無法溝通而又需要溝通。這些都因為一個基本道理：世上並無完美，而智慧只來自對缺陷的體認。這個世界是如此地糟，因為一半的人都是混蛋；它又如此美麗，因為一半的人都是君子。用我們的話來說，那就是說：如果 X 是常態，則 $-X$ 也是。

讀到這裡，你應該可以懂得，為甚麼太極圖有兩種讀法：白魚有黑眼睛，黑魚有白眼睛。但要這個道理用實在的言語、公式或者計算加以闡述，而讓河圖、洛書、易經這些東方的古早智慧不流於風水、命相、五行、八卦之學，從而能漂漂亮亮地對世事萬物的現象，在能解釋的時候解釋，在不知道的時候也知道並承認尚不能解釋，然後我們才能從玄學中昇華出來。

統計學是我們適當的搭配了我們對常態分布的了解（也就是說對於誤差及變異的了解）之後所衍生出來的歸納和演繹的方法。用統計建模的方法解釋世事萬物的現象，除了有「知之為知之，不知為不知」的誠懇之外，因為再加上對於不知部分的科學描述，而增加了分析的力道。而這在以前，是完全沒有辦法做的。

可以這樣說：發現並發展一套以常態分布為基石的分析方法，是二十世紀人類的重大成就。這套統計建模的方法真的有用嗎？我們用一些方法，並不一定知道它們是對的。有很大的一部分理由，反而是除此之外，並無別法。造物主是何等神奇，人類的知識，又是何等的淺薄無助。這一點點微光，在自然的神奇、甚至在人文的多變之下，既是極力的掙扎圖存，又是非常的力不從心。因此真正的智慧，反而來自謙卑：「所有的模型都是錯的，但有些是有用的。」說這樣話的人，應該是真正的智者：阿爾伯特、愛因斯坦。

九、統計是數據科學

統計是數據科學

趙民德，在人民大學的講稿一

前言

統計學的發展，是跟著數據的形態和問題的本質來改變的，不是因為我們會做它背後的數學而發展的。但在統計研究的工作的過程裡，不時會有一些還算是有趣的小數學問題出現。這是一個陷阱。不要捨本逐末，輕易地就掉在數學的陷阱裡。賀吉士（J. L. Hodges, Jr., 1922 - 2000）曾說過：不要因為（統計的）問題困難而去做它；也不要因為它難而不做。研究的選題，主要還是要看問題值不值得。

這是一個系列講稿的開場白。信息十分明白：統計雖然無可避免地用到很多數學，但不是數學。數學不夠好，做統計的工作，不論是研究還是實務，都很辛苦。但是數學太好，其實也有害。——每一個學門，都有她自己的味道。

統計學的味道，來自她的數據。因為這是她的本質。

因政令的需求而有統計

古早的統計我們不談¹，但世界上範圍最大的統計專業學社：國際統計學院（International Statistics Institute, ISI）就有超過一百年（1885 - ）的歷史。古早的統計來自施政（如抽稅和征兵）的需求，時至今日，政府統計仍然在各國的統計專業團體裡佔有舉足輕重的角色。因為它常川而定期地蒐集資訊，是任何有責任心的管理單位都不能忽視的工作。

政府的統計工作，是以蒐集、整理、發布為骨幹。它不止是在看全國，也要看年與年間的變化。所以橫截面（crosssectional）數據和時間序列（time series）數據，便成為主流。但因為這種類型的數據在型態上變化較小，且對於它們的處理方法，不論是做得好還是不好，都至少有五六十年以上歷史的延續發展，且都已在各國的政府統計中到位，在一般的統計研究中，較難有令人興奮的結果。

新的事物當然仍在開發中。當政府部門在大量使用電腦的時候²，當社會對隱私權日益重視的時候，當民眾對環境和生活品質要求日高的時候，政府統計的方法和態度，當然也會跟著調整。但是大體上，政府統計和經濟這個領域密切掛鉤，是大約不會變的。因為政府的目的是讓人民活得好，而經濟學用較全面的角度來關注這樣的問題。

農業實驗

一般以為，費雪（R. A. Fisher, 1890 - 1962）的工作，若不是打開了近代統計學的大門，也至少是具有繼往開來的氣勢和影響。他的書"On the

¹ 例如我們可以上溯到《管子》的《牧民篇》，這也許是中國最早談到統計的文獻。

² 最早使用電腦的政府單位，是美國普查局。

mathematical foundations of theoretical statistics" (1921) 雖談的是「數學基礎」，但是他工作的動機，卻是非常務實的農業實驗。真正的需求，不是來自小布爾喬亞式的「仕女品茶 (Lady tasting tea)³」的休閒題材，也不是來自「身高和體重的關係」這類的習題型的應用 (Galton 1877)，而是來自農業實驗：我們需要有效地安排諸多因子和它們的水準，以期能使得每一個觀測值都能發揮出極大的效用。這樣才能學到怎樣的實驗安排或肥料配方可以增加小麥的產量。這些是早年的 know-how。早先的科技，主要是在農業。這些工作的主題所包含的，不止是科學上的求真，也充滿了實務上、經濟上的意義。我們知道，歐美各國的工業，早期根本沒有競爭。所以早年還談不上統計在工業上的應用。

這樣的客觀條件，產生了在 Rothamsted Experimental Station⁴所發展出來的「實驗設計」。早期台灣的統計學，也以農業實驗為主（台大生統實驗室 1946）。早年台灣的經濟，靠農業很撐了一段時候。今天台大的農藝系仍然是台灣生物統計的重鎮。

實驗設計的主題，談的是如何有效地蒐集數據，以避免結果混淆和增加實驗的精度。科學實驗的目的是求得知識。現代的科技，近五十年來的進展，大概是超越了過去一百年前再往前上溯五千年間的總和。主要的理由來自人類開始懂得如何有系統地來由實驗一步步地向真理逼近。

敘述統計

較早的統計，大多是描述某一母體的統計學 (descriptive statistics)。時至今日，我們仍然在學：如何用少數幾個參數 (parameters) 來描述一個母體。「母體」的概念，也許是最要緊的統計想法。這是最早、最基礎的「模型」。進一步地，我們有所謂的「相關係數」。這才談到了兩個母體或者最簡單的直線迴歸。

從敘述到推論，中間僅隔了一線：但是古典和現代就在此分野。我們將問題中的數據架構在某一個機率模型裡，而忽然可用的解析工具就多了起來。我們學到的數理統計大概都在這樣的條件下誕生。它賦予古典的數據一個骨架，使得討論漸趨嚴謹精緻⁵，同時也蔓生了許多枝節，有時也不免將我們導入歧途。

在某種意味之下，數理統計的架構，提供給我們一個理想的想像空間。它既能概略地描述數據所給出一般印象，又提供我們能夠精密計算的可能。在精神上和實用上都是無可比擬並且很難拒絕的工具。統計學的基本理論架構——不論的假設檢定還是參數推估——是美麗而眩目的。但過度地強調細節，不免有如今日女仕們喜愛的名牌精品：雖然拳拳不空，但總覺得沒有真正到位。精雕細鑿之後的統計方法，似乎仍然只能解最初的問題。

基本問卷和迴歸

然後社會科學大行其道。或者，我們可以說：統計方法在社會科學裡大行其道。因為，以量化為主要手段的討論，是任何號稱是科學的學門最不能拒絕

³ 但這也許是「假設檢定」的濫觴。見 David Salsburg (2001)。

⁴ 這據說是世界上最早的農業實驗站，也是英國最重要的農業實驗站 (1843 -)。但他們不知道，康熙曾做過水稻的品種改良工作，而康熙末年 1722。

⁵ 這是學者們的最愛，因此生出許多事來。

的誘惑。量化需要數據，而問卷是社會科學中蒐集數據的基本方法。不能避免地，問卷裡會包含學研究者認為「可能會發生影響」的自變數。因此基本問卷的型態，幾乎是所有統計軟件的基本數據格式，而迴歸變成標準的最起碼的分析工具。這些都可以放在傳統型的 Fisher 架構上來討論。並使得一些統計名詞，例如 p -值， R^2 ，因為實在常常在相關的技術文件裡出現，便逐漸演變成爲現代科學語言的一部分。

承認隨機性

統計學發展主要的影響，是科學界開始承認隨機性的存在，並務實地和「誤差」共同生活。我們不止有了風險的概念，並且還能進一步將它量化。

將這個概念推到商業的極致，便成爲今日的保險業。

承認隨機性並試圖針對它加以計算，是新的想法。科學不能避免量測，而量測不免誤差。我們經由誤差來了解模型，經由變異來了解風險：這背後是我們對於正態分布性質的澈底追尋。

這樣的進步是劃時代的。因此證券的投資比例（portfolio），可以用一個多維的 mean-variance 架構來加以表現，這是所謂的資產定價模型（capital asset pricing model, Sharpe (1964)）；在另一面，衍生性金融商品的價格，必然地和選擇權的對象的價格的變異有關。這是所謂的 Black-Scholes (1973) 模型。這些都是精彩的論證。

品質管制的興起

戰爭除了破壞之外，還大量地刺激了消費和需求。二次大戰（1931 - 45）使得大規模而快速的工業生產，不只成爲必要，且延續到所有和軍品相關的工業。這時便自然地產生了諸如品質管制（quality control, QC），運籌學

（operation research, OR）這些學問。這裡面，QC 的部分，並沒有太多的正規統計推論。但是卻是有聲有色的應用。而 OR 卻較偏向最優化的數學工作。但這兩者，不論是規模和影響都十分巨大。比較早期的 OR 專題之一是排隊的理論。這也是來自工業上的需求：電話網路的安排（趙民德 1975）。

早先關於 QC 的數據，是生產線上自然產生的（production line data），我們只要費心加以測量便好。QC 的技術，著重在快速有效，並且還要有在第一線的人員就能明白的意義。雖然，在技術上，如果我們能充分懂得二項和正態分布，早年的 QC 大概就不出這樣的範圍。但這不是我們的重點。我們要強調的是：數學上困難度和統計學是否有重要的影響無關。課題本身的專業意義，要重要得多。

OR 的另一個方向，又漸漸演變而爲管理——數量管理。這些不全是我們所熟知的統計學，但是不論是簡易的「品管七大手法」、TQM 或者目前還在大行其道的 6- σ 系統，其基本可以實際運作的部分，仍然是最起碼的統計。統計方法不斷以不同的型態出現，例如在資料倉儲（data warehouse）裡的核心——是一個稱爲 cube 的數據結構，而這只是我們常用的多維列聯表。

到目前為止，數據的形式都還是完整的。當我們說「看到 X」，意思是說，我們可以完全看到 X。但是，漸漸就出現了「只看見一部分 X」的情形。我們下面所談的就是這樣的情形。

設限資料

我們先是知道資料有用，並且盡力蒐尋，努力將其中的資訊擷取展示。早先的資料都是完整的資料，對於不完整的資料的分析，一方面需求不高，另一方面我們的能力有限（例如五零年代的電腦，不能和現代相比）。因此早年的發展，除了一二個「能算出解析解」的特例外。結果較少。

不完整的資料，來自實務。對於純粹的「遺失」，我們不妨將它們簡單地看成「根本沒有蒐集」。例如我們原想取 100 個數據，但其中某一個沒有取到，那麼我們就不妨將數據看成一組有 99 個數字的隨機樣本。——這個想法，基本上是對的！但是，若是我們對某些數字「只看到一部分」時，又應該如何處理⁶？

「只看到一部分」，可能是時間不夠，可能是目標已達。這些都是簡單的「設限數據」(censored data)。例如對於癌症的臨床研究，通常不會超過五年。時間到了，病人還健在，我們只知道他「存活了至少五年」，但看不到他會存活多久。這是第一類的設限。這在工業可靠度的研究中也時常發生：現在的工業產品愈做愈好，在 100 個測試的產品裡，或者只有 70 個或更少的真正是損壞的。那麼，餘下的 30 個未壞的產品，難道不包含可用的資訊？

在生物臨床實驗裡，不完整的資料還會以其他形式出現。例如病人可能是死於車禍。由於「生命分布」愈來愈長，關於正確處理種種設限數據的技術，才會風起雲湧地被發展出來。這背後的力量是需求：製藥的確認以及醫學上的求真。當然它們都需要相當的數學技巧和深度，但那不是重點：重點是，有太多這一類的數據需要我們去分析。所以才會發展出新的統計方法。

加速壽命測試

對於人體，我們不能對他做出過於冒犯的事。但是，對於電子元件，我們大可以加上一些條件，使它們會「壞得快些」。例如增加測試時的溫度、電流強度等。這在可靠度的研究和實務裡，叫做「加速壽命測試」，是臨床醫學裡沒有一項。

對於某些產品，它基本上不會壞，那麼又該如何去做？例如發光二極體(LED)的壽命，理論上是十萬小時。沒有一個工廠會等那麼久，一萬小時都太長。這時，我們就要考慮「衰變分析」，例如每周測一次它的光度，看光度減弱的程度來研判估算該產品的可靠度。

高可靠度分析，來自精密工業的要求。但是，科技又讓我們有其它不同型態的數據。在在需要統計學家的投入。

數字型態以外的數據

數字型態以外的數據，比如晶元圖(wafer map)，生物晶片(micro array)，衛星遙測資料，煉鋼爐或者某一個化學反應槽的內部反應。每一種資料，都有它蒐集的目的，也有它們想問的問題。但它們多半不是 H_0 vs. H_1 的教本型態，我們也最好不要先三不管將它們化為假設檢定或者信賴區間。

但是統計學的基本道理，也許會有所修正，但不會有太大的改變。這都拜

⁶ 簡單的方法是將它想成遺失。很多人現在仍是這樣做，但這樣會數據裡的損失資訊。

近四五十年統計學者對於統計學基本認知的努力。但問題會一變再變，數據也會有不同的形式，只有不斷地從新型數據裡來一再深探，統計學才能夠真的成長。

很多統計學者是學數學出身。數學裡有很多有名的大猜測（conjecture）。任何一個小數學家，若能證明其中之一（甚或一半），都會立刻成名。這是數學家看重的。但是，統計學裡沒有甚麼重大的猜測。你能證固好，根本不去想它們也不會被別人說功力不夠。因為若是我們去看真的問題和真的數據，我們根本沒有時間去玩數學遊戲。

統計科學還是統計工業

前面我們所提到的多半是統計學，或者更進一步的統計科學。但是，科學還是有一點不夠，我想看到的是統計工業。

工業意味著不斷改進的、有市場的產品和日益增加的就業人口。如果統計只是在大學裡或者研究單位裡面存在的話，那麼個行業的蕭條，也是可以預期的。

工業的另一個意涵是不再有單兵作戰。關張趙之流的勇將，還是需要大部隊的支援。工業化的結果，除了世俗化之外，還暗示著有更高的附加價值。有了小兵，才襯得出大將的不凡，而小兵也需要養家活口，盡他們的社會責任。

我曾在化廠裡看過滿牆的螢幕：每一個都在監測某一重要生產環節的狀況和反應。我聽過虛擬的晶圓工廠，我也看到企業裡開始有「資料價值發展部」的編制。教育測量服務（Education Testing Service, ETS）所做的，除了 TOFEL 之外，還有不斷改進中的測驗技術。這是將 item response theory (IRT) 充分發揮的工業。SRS 和 NORS 基本上都是企業化之後的調查機構。而 Morgan-Stanley 難道不是在蒐集並發布數據？那些作債信評等的大公司，處理的不是數據？甚至於華爾街日報也做的是類似的工作。——他們只是將某些統計工作專業化，不斷地加上新的價值罷了。

結語

魚要有水，統計要數據。水若污染，魚會死，因為它改變自己不夠快。統計要能存活，或者，一個統計人想要存活，不要只守著一畝三分地，只做自己或者別人方法論的推廣，更不能甚麼問題都套上同樣的三斧頭。要不斷做新的投資。

自從人類懂得結繩記事，我們就開始有數據。科學的發展，使得現在更是一個充滿數據的時代。蒐集它們都有一點目的，有些數據來之不易，成本甚高（例如航太遙測的數據），它們背後的目的更大。數據是時代的大河，千里而來，出海而去，我們知道泛濫後的土更為肥沃。

統計學的長遠發展，要建立在這樣的基礎上。

參考資料

趙民德 (1975). 從鼠牙談起——五花八門的電話研究工作 科學月刊全文資料庫：<http://140.111.102.168/science/content/1975/00110071/0008.htm>

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637-654.

Galton, F. (1877). Typical laws of heredity. *Nature* 15, 495-533.

Salsburg, D. (2001). *The Lady Tasting Tea*. W. H. Freedman and Company.

Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 425-442.

file=D:/my stat articles/paper-7.ctx, printed April 26, 2004

十、關於統計論文的撰寫

關於統計論文的撰寫

趙民德

中央研究院統計科學研究所

2006/9

在人民大學

前言

這本小冊主要討論的是如何撰寫統計學的論文。雖然對於一般的科學論文有所著墨，但是以統計學為主，而且是以統計學的方法論為主。

撰寫論文其實只是末節，研究的本身才是本質。「為學問而研究」和「為發表而研究」是兩個層面。我退休之後有時會選一兩篇經典論文細讀，才發現有好些以前沒有唸懂的地方。甚至可以說，如果當年真唸懂了，有些我自己的研究工作，可以少走不少冤路。道理其實很簡單：功力不足便要上路，總免不了跌跌撞撞。

但是人類的進步，都不是來自四平八穩的工作。如果都要等所有的預備工作都好了才開始寫論文，恐怕一輩子都不夠做準備的。

第一部分：心理建設

在大學任教，以前除了到時候去上課，學期完了給成績，對於是否要寫幾篇論文，多半是自願的事。現在時代一變，變成了「不發表就走路 (publish or perish)」，因此，發表論文，似乎變成比好好教書更重要的工作。

這本小冊，表面上討論的是如何撰寫可以發表的論文，但我想深一點談關於論文寫作的種種，因為學術上的功力是本質，而如何撰寫乃是末節。

沒有研究工作就沒有論文

發表論文的背後基礎是研究工作。做研究工作有如尋礦：你沿著一條或者幾條線索，用某一種工具一點一點地挖下去，希望看到有價值的東面。這裡「有價值」是比較抽象的。開礦挖到了金子、石油就算成功，最多是蘊藏量的差別罷了。但是對研究工作而言，「有價值」指的是「新的」，「能承先啓後的」，「替別人開路的」。這裡，「新」是必要條件。新結果在理論上指的是至今人類都還沒有做過，並且是 non-trivial 的結果。

再好的結果，如果是幾年前被其他人先一步發表了，除非你能夠再弄一些新見解，那麼你就差不多等於做了虛功。你可以有一點虛榮心的滿足：我所做的研究和某某人一樣好。落後別人之後，如果你一再試投也許還可以在較弱的期刊發表出來：希望他們的評審者學問不深，看不出來。

但這不是正途。發表論文並不是沒有邪魔外道但仍勉強合乎學術倫理的辦法。但這些方法既弄不長，也搞不大。靠這一套辦法（如果你夠小心的話）可以混一口飯吃，並且可以混得不壞甚至於可以小有名氣，但這些之上，如果還加上「我有一點真功夫」，那麼這些所謂的邪魔外道的本事，就可以變得名正言順。

這本小冊，主要還是先鼓勵大家先有些真功夫再玩別的遊戲。沒有實力，甚麼都是假的。

為甚麼要鼓勵學術研究？

主要的道理：科學¹上的實力，是一個國家最基本的實力。論文發表得多，表示有很多人在做科研，表示這個國家有學問有見識的人多，代表這些可以反映在社會和諧進步，人民生活美好以及船堅炮利。

但這雖是一般領導的簡單想法，卻和真正的科學本質是有差異的。讓我們做好研究的基本力量，雖然很大程度上包括了政府和社會的鼓勵和壓力，但能夠讓第一流人才半夜不睡搞科研的基本力量，並不是這些。這類第一流人才，如果他要在政府的獎勵條件之下優化他的科研成績，根本用不著這樣三更燈火五更雞那樣的拼命。

深入推動研究的力量

真正的工作是好玩的，是有癮的。是由興趣，好奇心，探險的精神這類事物後面

¹泛指較軟性的社會人文科學和物理化學這類傳統的硬科學。

推動，才能讓一個智商很高的人²，圍繞著一兩個問題的焦點，東翻西找，千尋萬覓地想了解真相。好的科學研究結果，多半來自探索未知。

我們引用一段散文：

在美國找一個書店是不容易的事，找到一個有點正式的書的書店，更難。而我信步所之，在這麼個小街，發現這麼多可愛的書。眼前好像有一片眩目的光芒。掏出一把在倫敦機場換來，還不會用的錢，讓店家挑了兩先令去，帶回一本《新科學家》來。

三翻兩翻，即看到很熟悉的幾張畫。仔細一瞧，這些書全是中國的東西：一張是一六零一年湖北的鐵塔；一是一六二一年射的火箭；一是一六一零年河北的拱橋；最好玩的是一張木刻，是用河水推磨，用風箱吹起旺火。

這是一個叫做陳之藩的學者，當年從美國到了英國的劍橋去訪問研究，在那樣的環境下深思反省之後所寫的一系列散文中的一段。那本書叫「劍河倒影」，曾經給台灣的學子深刻的影響。陳先生是中文素養極好，而科研也卓有所成的學者。上面的一段話，來自他讀了丹尼約瑟的《東西方的科學與社會》（此人在 1947-1964 間的論文集）這本書之後的反思。這本書講的是中西科學的比較，中國科學對西方的影響，以及中國社會與中國科學的關係等。

作者發現很多科學來自中國，而一個自然的問題是：為甚麼那麼多的發明，卻沒有導出像歐洲近五百年的科學發展？

這樣的問題，我們中國人何以沒有被認真地問過？何以沒有被好好回答？我們時常說「XX 中國古已有之」，或者搞出一些四不像的如「中學為體，西學為用」的道理，更簡單或者更浮面的做法就像李鴻章那樣的用銀子去買外國兵艦來建北洋海軍。但是，近代中國的積弱卻是一個事實，而這個事實在最近 20 年才開始有所改變。

丹尼約瑟的結論，主要在中國科學在整個發展中主要是為了「實用」，而歐洲近五百年的科學發展，主要是為了「好奇」。

中國人太務實了。如果歷代將「奇技淫巧」當作科學的測驗項目，相信我們的科學會全然不同。但這樣做，也許可以產生一些第一流的工程師，但還是產生不出學術上的能引領一代風騷的大師，就像歷代的科舉狀元，他們也許能留下一兩首詩，但對於學術的貢獻其實並不大一樣。

丹尼約瑟以半生的時間跑完了中國，又淹在劍橋的書海裡，去發掘中國的科學史，這裡除了「好奇」，還能說出其它原因嗎？反過來看，我們能不能找到一個中國的丹尼約瑟，以半生的時間，淹在南港的書庫裡去研究歐洲的科學史來解答「歐洲近五百年的科學發展，主要是為了好奇」這個假設是對還是錯³？

陳之藩最後說：在這種笨人不能產生之前，我們所謂的科學，還是抄襲的，短見的，實用的。也就是說，真正的科學，是不會產生的。

類似的反省，在時下的網路小說裡也看出一部分來。這類的小說，算是「架空歷史」

²我們不敢說研究工作者智商一定高，但這一批人的確是較一般人通過較多的考試，所以至少是一批國家菁英。

³這一段是陳先生的原文。

類。多半說的是某一個知青，忽然回到過去，如何用他們的現代知識來改變古中國的命運。但他們的著眼，仍在科學的應用和制度的改革上，對於「好奇」的概念，沒有著墨。

研究的目的

好奇和實用：這兩點我們不需再說。

明確的了解：把原始論文說得更清楚，以便後人更易了解。例如 Taylor 定理的原始證明是很長的，現在任何一本微積分大概都用不到一頁。又如以前號稱世上只有七個人讀得懂相對論，但現在任何一個物理系的畢業生都能懂。

確認：做科研相當於人與天爭，因此要步步為營，以便確認之後的結果後人敢大膽地用。

國家鼓勵研究的目的

這是使國力強大的投資。智識就是力量，就是財富。這表現在戰爭中尤為突出。雖然，搞科研的人士，一般都傾向於反戰。他們想看到的是贏在實利：國民所得逐年增加，Gini index 逐年減少。

使國力看起來強大：這是很多學術官員能夠做的。在全國的層面上，他們只能做一些以論文的數量和質量為標準的獎勵辦法，但是上有政策下有對策，早晚這類獎勵發表的政策和計算論文的方式都會被某些人破解。

關於 SCI/SSCI

亞洲各國的學術官員對於 SCI/SSCI 有一種莫名其妙的信任。但 SCI/SSCI 的唯一道理，是「它不由國內的教授控制」。但這最多只能拿來做過渡的手段來用，不能長期倚靠。學術最終還是希望在國內有一套有聲望及傳統的學術評鑑機制。一個大國，必須要能做到「自己說了算」。雖然這一點並不容易，全得靠硬功夫和好的學術倫理和制度。

下面的信是 1962 年陳之藩寫給胡適的：

告訴您一個好消息：我又有一篇論文，就在 IRS 印出來了，...這篇文章很短，但還乾淨，題我叫「A new approach to Fourier coefficient evaluation」，您大概數學都忘光了。這是數學上的大問題，我提出個圖解的方法來。有個數學家說，如果世界上的書全要燒光，該留 Fourier series。我的貢獻並不大，但寫得很俐落。這是照著您在東廠胡同教我的八字秘訣：——「開門見山，水清見底。」開門見山不難，少廢話就行了。讓水清見底是談何容易的事。

有沒有現代的年青學者，給他的博導寫類似的信？這是我們傳統裡有味道的部分。如果他寫成這樣：

告訴您一個好消息：我又有一篇論文，就在 *** 印出來了，...這個是 SCI 期刊，以 Impact factor 而論，在同類期刊中排名第四 ...

這也許還有一些尊師重道的部分，但是深度就差了。但這是現在的遊戲規則⁴。寫到這裡，真免不了有很多無奈。

心理建設

這一部分的目的，主要是讓大家對於「寫論文」一事有一些心理建設：有些事情是需要了解一下的。

1. 能刊載的論文未必好：能刊載只表示通過某學術期刊的評審而已。
2. 被拒絕的論文未必糟：有幾類論文容易被拒：相當糟的論文，不討喜的論文，擋著別人路的論文。

SCI/SSCI 等只是參考：它主要是以某 data base 上的紀錄來決定是否收入，因此較不含個人的喜惡。觀念上我們別將它們看得那麼大：基本上 impact factor 只是 popularity index。也別將這些看得那麼小：意思是說，亞洲各國的科技官員也並非那樣的不了解狀況。對於 $(100 - x)\%$ 的教授而言，登出一篇 SCI 論文，也並不容易——寫一篇論文的基本功夫你還是要做。對政府的鼓勵和獎助，年青時不妨積極些，年紀大了，就該用平常心看待。

這一切全在您把自己看成是哪一類學者。這句話說得比較有鼓勵性。但我的意思是：做一個能讓人尊敬的學者，其實並不是遙不可及的事。和最好的學者相比，一個實在的博士，在專業上也不過是相差在兩三本真正好書和十幾篇真正弄懂的經典論文。如果客觀環境略好一點，有一兩年的「折節讀書」也就趕上去了。但是，維持在第一線，也需要用心用力。

原則上，寫論文的技術好學，但實質的部分則需要功力。而功力指的是學養、品味、苦工、眼界和心胸。寫論文並不是做燒餅油條，甚至不也是製造 IC 晶片。在學術圈，「人民的眼睛是雪亮的」。長期下來，只有真正的學者風骨才能得到尊敬。

功力的來源：這包括了先天的優勢，自身的努力，師門的教導和保護，工作團隊的合作，秘笈（一個好圖書館），諍友和好學生。

好文章

好文章有很多看法。但是有一點是公認的：前 $x\%$ 的文章（而 x 的值蠻小），才叫好文章。因此，大部分的作者，大多數的時候，都只是在做不怎麼好的文章。這裡並沒有太多的不滿，因為沒有那 $(100 - x)\%$ 的普通，怎顯得出那 $x\%$ 的不同？因為國家大力要求論文的發表， x 的值只會愈來愈小，因為分母在持續增加。可以這樣說：論文超過某數的作者，當然是會有好作品的，但多數作品的也不會太好。

別好高騖遠，儘找自己做不出，別人也做不出的題目；別急功近利，儘找一定可以發表的題目；這兩者都很難增加您的學術地位。前者是因為多半您也做不出來；後者則是別人也看得出來您沒有多少貢獻。對您的學術生涯規劃，並不理想。

較佳的選擇

⁴這是以西方的價值為價值的規則。一個大國應該有她自己的志氣。

做比自己能力略難一點的題目；這樣才有一點挑戰性，別人也不會覺得您總是柿子挑軟的來捏。長久下來，會對您產生一定的敬意，自己的滿足感也有一些。

專題的範圍，不要太窄，最好要有中長期的布局；這樣，十年下來，您可以有能「成一家之言」的機會。

不要東打一拳，西踢一脚。這種亂槍揚打鳥的方法，投資報酬率太低。

September 4, 2006

第二部分：好論文的特質

甚麼是好論文？這是較難的問題，但卻是要緊的問題。學術期刊都在挑好論文，因此較好的論文才有較佳的機會發表，即做是純粹爲了發表，寫好論文也是重要的考量。

寫作是有技巧的，但我們還是認爲：內容第一，包裝第二。事實上，能做出好內容的學者，寫作的本事也不會太差。因爲他們既然能夠在千絲萬縷的迷宮裡發現出前人未知的道理，這些人對於邏輯推理以及清楚表達的能力，應該都是沒有問題的。的確，有的時候，某些結果是經由作者經由他獨立思考方法東拐西繞之後才得到的，如果沒有經過清楚的沉澱，寫出的論文只沿著作者原來幽暗的思路而發展，其他的讀者未必能懂。有時作者認爲自然的事物，讀者未必也認爲自然。好的結果未必好懂，但是內容，仍然只有內容，才是論文的本質。

有人說，好論文如美女，看到了就知道。錯！美女一看便知；好論文只有較高手的同行才知道。因爲學術論文這檔子事不是普羅文學，而是一種精緻文化。

低段的棋手，往往不能看懂高段棋手的布局和收官——這需要足夠的背景資料的瞭解。學術論文絕不是通俗文化，它是非常講究的、針對一小部分有足夠知識的讀者所作的展示。理論上，一篇論文應包含所有能讓讀者讀懂該文的資訊。學術論文裡不見得沒有錯，但那都該是「誠實的錯 (honest mistake)」，因此絕不該有故意的錯。像金庸小說裡郭靖故意搞錯寫一部九陰真經去胡弄歐陽鋒的事，在學術論文裡就是犯了大忌的¹。

讓我們來隨便看一看甚麼叫做學術上的好。我們說「隨便」，因爲自己見識也是有限，並且主觀也強，是不可能將所有關於好論文的條件都一一列舉的。讓我們從最簡單的一個公式說起。下面的一個公式是號稱是世上最漂亮的公式：

$$e^{i\pi} = -1 \quad (1)$$

這裡面有， $e = 2.71828\dots$ 和 $\pi = 3.14159\dots$ 這兩個各自合乎天人之道的「超越數」，再加上 $i = \sqrt{-1}$ 這樣的一個虛數，這三件東西怎麼會巧妙地連在一起？

科學家心中的審美觀是非常主觀的。我講不出十分有說服力的道理。但是， $e = 2.71828\dots$ 是來自

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$$

的極限，而這成就了「複利」的概念：生長的速度，和自身的大小成正比。這正是自然界「生生不息」的描述。世間有那麼多「exponential growth」的模型，並不是偶然。而另一面， $\pi = 3.14159\dots$ 來自圓周率，從這裡造就了多少三角函數這種周期函數。是的，「周期」是我們的關鍵字，因此 π 和我們所知的「循環不息」密切相連。而公式 (1) 巧妙地將這兩種合乎天人之道的事物連在一起。

我不是在講《易經》。但公式 (1) 是明明白白的可以用最嚴謹的數學講清楚的事！

甚麼是「好結果」？首先的要求是「並非無足輕重 (non-trivial)」。大體來說，凡是能丟給學生做，並且你也相信他或她大概會做得出來的問題，大概就算是 trivial。大部分

¹情治人員可以在某磁片上加一點或減一點去胡弄對方的間諜，但是在學術期刊上公開發表的論文就不可以這樣做。

博導給研究生的題目，雖然不一定算是 trivial，但也不會相差太多。原因無他，如果不是長期浸淫於某一專題並有相當的功力，好題目一般並不容易找到。如果有，博導自己為甚麼不做？

能問好問題，並解之

年青時常聽到一些老數學家說：「證明定理不難，難在寫出定理的敘述」。現在年紀大了，有時覺得寫出定理也不那麼難。玄幻小說裡常提到，修真的人功力一深，便漸有預知的能力。做科研久了，對於自己浸沉已久的專題，甚麼東西會成立，甚麼事情大概不會對，多多少少是可以猜出來的。

所謂的好問題，是我們一般猜不出答案的。但當你一步步把它梳理清楚之後，它的答案又明白不過。這裡，你的學術貢獻有二：(1)，問出好問題；(2) 解決它。

古典的統計方法 (例如在如 Kendall, Stuart and Ord² 這樣的經典教本中所列舉的方法) 多假設

$$X_1, X_2, \dots, X_n$$

這組資料是給你分析的。基本的想法是別人已做了實驗，取得數據，現在請統計學者去分析。這類的架構，在五六十多年前是標準的。在數學上，因為

$$(X_1, X_2, \dots, X_n) \in R^n$$

其中的樣本數 n 固定，是一個有限維 (finite dimension) 的問題。所有的統計推論，是根據這組數據的聯合機率密度函數

$$f(x_1, x_2, \dots, x_n | \theta)$$

而發展的。這裡的要點是： n 是固定的。

要打破這樣的架構並不容易。在這架構之下，有些事情是做不出來的。例如即使在

$$X_1, X_2, \dots, X_n \sim \text{iid } N(\mu, \sigma^2)$$

這樣的最標準的模型下，要算出一個關於 μ 的長度為給定值的一個信賴區間，也做不到。比如說，教本的做法是利用

$$\frac{\bar{X} - \mu}{s} \sim t_{n-1}$$

來造出

$$P[\bar{X} - st_{n-1, 1-\alpha/2} < \mu < \bar{X} + st_{n-1, 1-\alpha/2}] = 1 - \alpha$$

其中，這信賴區間的長度是

$$L = 2st_{n-1, 1-\alpha/2}$$

而因為 $s^2 = \sum_i (X_i - \bar{X})^2 / (n-1)$ 是一個隨機量，不會等於任何給定的常數。

且不問為甚麼我們有興趣去找一個「有固定長度的信賴區間」，至少在幾十年前，這是個大家都知道的問題。若是誰能解出來，總有一點學術上的功勞³。這裡面有兩條路可走，第一是設法證明這問題做不出來；第二是換一個條件來把這個問題做出來⁴。

²這是三冊的 The Advanced Theory of Statistics, 初版在 1961 年。

³五六十年的數理統計和數學非常掛鉤，這類蠻數學的命題，還是有一點市場的。

⁴這裡面有一個次序問題。先做出第一點再做出第二點是高手；雖然多半的人是先做第二點。

證明這件事不可能是可能的。這是 C. Stein 的工作⁵。這是個好結果，因此可以刊載在好期刊上。但是，證明「不能做」某件事，因為別的數學家老早就證明了「只用直尺和圓規不能做出三等分一個任意角」，Stein 的結果，就不能那樣的劃時代了。因為在三等分角的問題被解決以前，人們根本不知道「某命題不可能做出」這件事居然是可以證明的！

既然可以證明在樣本數 n 為固定時做不出來，那麼在 n 不預先固定的時候呢？Stein 的解法大概是這樣的：先取 n_1 個樣本，做一些計算，再根據這 n_1 個樣本所算出來的結果，來判斷是否我們還需要再取 n_2 個樣本——視情況而定。在這樣的安排之下，Stein 可以造出一個關於 μ 的長度為固定長的信賴區間。注意到，他所用的樣本數不固定：有時是 n_1 ，有時是 $n_1 + n_2$ 。

針對同一個問題，Stein 就至少有兩篇好論文出來。但這問題並沒有結束：Stein 的信賴區間是否可以做得短一點？是否已做到最短？是否可以做到 normal 之外的分布...

類似的問題可以問得更深入。隨著問題的深入，題目也愈變愈窄。解法和所需要的數學也愈來愈難...。這在序列分析 (sequential analysis) 裡，最後成了一個蠻特殊的專題。

Stein 的工作，告訴我們在樣本數不固定時，可以做出一些樣本數預先已固定時我們做不出的事。這個結果在概念上給我們的啟發，遠大於技術面的貢獻。而其它諸子的工作，就多半只剩下數學上的困難，就味道來講，就不免要弱一些了⁶。

我們再舉一個例子。在

$$\int f(x)dG(x)$$

這樣的寫法裡，我們基本上要要求 $G(x)$ 是一個平滑函數。雖然我們可以對 $G(x)$ 的性質降低一點標準，例如在 Riemann 積分中，如果 $G(x)$ 是一段一段連續的就好。而在 Lebesgue-Stieltjes 積分中，我們只要求 $G(x)$ 可以用來定義某一測度。但這樣的條件不能降得太多。因為 $dG(x)$ 這一個符號，講的就是，如果將 $G(x)$ 一小段一小段來看（不論是對 x 軸還是對 y 軸），因此對它的某些連續性，總是需要的。

但是如果 $G(x) = B(x) =$ a Brownian motion，這個積分的問題便完全不一樣。這是因為除了它的隨機性以外， $B(x)$ 還是一個到處連續，但無處可微分的函數。

這時，要能夠清楚的說明白甚麼叫做

$$\int f(x)dB(x)$$

就是好結果 (K. Ito⁷)。因為要做到這一點，你必須跳出原有的對於積分的概念那類的框框，而重新看到並梳理出一些全新的東西。有了重新的定義，才有所謂的「隨機積分」這回事。

抓住問題的本質並說明

衍生性商品 (derivatives)，指的是證券的買權或者賣權那一類的未來合約。這種合約是可以交易的，而它的交易價通常是針對某一種上市股票的價格的變動而變動。例如 X

⁵感謝張源俊教授告訴我這件事。

⁶例如，在序列分析裡，因為數據是 $(X_1, X_2, \dots) \in R^\infty$ ，問題的數學結構要難得多。

⁷原始的論文大概是 1946 年。當然，現在多半的關於 stochastic integration 的書都會詳細介紹。例如 J. M. Steele (2001). Stochastic Calculus and Financial Applications, Springer, 第六章。

公司的股價現在是 30 元，三個用後，我是否可以用每股 35 元的價格來認購？如果可以，我現在要用多少錢來購買這樣的權利？

衍生性商品的市場值，現在已大於股市。是目前財務金融市場中最要緊的商品。那麼，它的公平價格，應該放在哪兒？

這個問題在 1973 年⁸ 才有重要的突破。甚麼叫做公平定價？一個重要的想法是：這個價格要能使得交易雙方都沒有完全不冒任何風險而輕鬆獲利的機會。在 Black 和 Scholes 的論文裡，他們將公平這個觀念，用「有效市場中無套利」這個基本條件來抓緊在一起。而在這樣的條件下用精準的數學來導出所謂的 Black-Scholes 公式。因為他們清楚地將問題的本質抓住，雖然 Black 和 Scholes 所做的是最基本的 European option 的定價，但這一點突破，卻是近代數理財務學中最關鍵的步驟之一。近十年來，它都是最火的學科，因為許多其它的較複雜的衍生性商品的合理定價，基本上也還是要靠類似的想法來推導。

我們再換一個問題，問一下「甚麼叫做 information」？

這個問題也不簡單，但我們不能平空來問這樣的問題。一個有效的方法是，先找到一個有用的架構，然後再試著把這件事說明白。

不同的架構下，information 有不同的意義。例如在統計推論的架構下，information 一詞，指的是 Fisher's information。它的基本想法是，數據是用來做統計推論的，而未知參數的估計，是統計推論的最基本的部分。因此，Fisher's information 所看準的問題是在問：對於未知參數的估計而言，我們最好能做到甚麼程度？

但對於通訊理論而言，問題的焦點就不在於參數的估計，而在資訊的傳達：一連串訊號從 A 點出發，經過一個介面，再傳到 B 點，關注的是：有多少東西送進了 A，又有多少東西由 B 傳了出來？

為了清楚地說明這件事，我們先得問「在甚麼情況下，通訊介面其實沒有傳遞任何訊息」——換句話說，我們先得知道甚麼是 0，然後才能知道甚麼是 1。

至於如何說明何謂沒有 information，一個比較合理的想法是先列舉幾個「不含 information」的合理條件出來。而在這幾個條件下導出 entropy 的觀點。

以上基本上是 Shannon (1948) 的結果⁹，這在通訊理論上是重要的一筆，甚至於發展出一個專屬學術期刊來¹⁰。

客觀條件成熟時的總結

人的想和和基本解析能力，其實是差不多的。如果某一類問題經常困擾著我們，那麼自然便會有許多不錯的學者去鑽研它。他們的解法，也往往有類似的想法出現。但每一個人的出發點不同，所面臨的數據結構也不同，因此看到的問題和解法也會有些共同點，也有些差異。時間一長，功力深厚的有識之士，便會看出這類問題的真正癥節，從而能做

⁸F. Black and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637-654. 多半的關於數理財務學的教本都會提到。

⁹C. E. Shannon (1948). A mathematical theory of communication, *Bell System Technical Journal* 27, 379-423, 623-656.

¹⁰IEEE Transaction of Information Theory 創辦於 1953，是 IEEE 的招牌期刊之一。

一個清清楚楚的整理。這類的論文，是一種「集其大成」型的論文。

例如 Dempster 等¹¹ (1977) 在已有很多論文討論不完全資料如何處理後，才推出 EM-Algorithm，將計算 MLE 的手段，歸結到兩個步驟。因為這個集大成的工作，涵蓋面非常的廣，幾乎將所有的關於不完整資料的處理方式，一網打盡，從此變成經典論文。

另一個在客觀條件成熟時才引起廣大迴響的論文是關於自助重抽法 (bootstrap method) 的，我們知道，一般的統計推論問題已基本上可以歸納到「求出某統計量的分布」。而古典的統計方法是利用解析上的功夫 (包括大樣本理論) 來計算該統計量的分布。但解析的方法有限，當所有的解析方法都已用盡時，我們還有甚麼其它的方法？

當電腦計算變得容易時，當幾乎人人都有一台 PC 時，當統計軟體變得普遍時，這時總結出一套「如何利用模擬以計算適當統計量的分布」這便是自助重抽法，見 Efron (1977)¹²，這也是一篇經典論文。

新技巧的引入

有用的新技巧的引入，也是好論文的條件之一。例如 fast Fourier transform (FFT)，可以將一個需要 $O(n^2)$ 次的計算，簡化成 $O(n \log n)$ 。又如 Markov chain Monte Carlo (MCMC) 的想法，基本上將 Bayesian 方法的單門——後驗分布多半算不出來——加以解決。這類的結果，當然會引起廣泛的重視。

當有一大類問題需要解答時

設限數據 (censored data) 是近代常見的數據，這尤其對於描述壽命長短的問題，經常出現。例如 Kaplan-Meier estimate，這是對於一組設限資料如何找出合理的對 cdf 的估計量的論文。見 Kaplan and Meier (1958)¹³。因為這類問題實在遇到得太多，而且這也是很多關於如何處理設限數據的基本論文，它差不多是世界上被引用次數最多的論文之一。

比估計 cdf 更進一步，如何對於設限數據來做迴歸分析，那當然是更會讓人用到的結果 (Cox 1972)¹⁴。如果數據中還有一部分相依，類似的迴歸又要如何去做 (Liang and Zeger, 1986)¹⁵？這類的結果，是因為太多組數據是這種形式而變得重要。

這類好文章的來處

這類所謂的好研究成果是如何來的？最自然的是因為功力到了自然形成。當你在某一個行業久了，你自然知道甚麼是重要的問題，甚麼是用一點苦功就能做但做出來也功勞不大的問題。因為為了要發表，你需要做新的東西。如果你對當前的一般行情都摸不清楚，你能玩這類遊戲的資格還淺。學術發表，就是要走在別人前面。

¹¹A. P. Dempster, N. M. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data via EM algorithm, *J. of Royall Statistical Soc. B*, **39**, 1-21.

¹²B. Efron (1977). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, —bf 7, 11-26.

¹³E. L. Kaplan and P. Meier (1958). Non-parametric estimation from incomplete observations. *J. of the American Statistical Association* **53**, 457-481.

¹⁴D. R. Cox (1972). Regression models and life-taaables. *JRSSB* **34**, 187-201.

¹⁵K. Y. Liang and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

我們常說，文章本天成，妙手偶得之。這一點，對於學術論文而言，往往不是這樣。一般來說，妙手很難偶得。基本上，即使你已到了作戰的第一線，你還是不能太鬆懈。因為和你競爭的對手都沒有鬆懈。

王國維在《人間詞話》中提出境界一說。謂

古今之成大事業、大學問者，必經過三種之境界。昨夜西風凋碧樹，獨上高樓，望盡天涯路。此第一境也。衣帶漸寬終不悔，為伊消得人憔悴。此第二境也。眾裏尋他千百度，驀然回首，那人卻在，燈火闌珊處。此第三境也。此等語皆非大詞人不能道。然遽以此意解釋諸詞，恐為晏歐諸公所不許也。

有人說做研究亦然。我比較相信前面兩個境界，那指的是苦讀和迷惘。至於頓悟卻是少，畢竟研究工作是靠邏輯、實驗和計算哪。

這類所謂的好文章很難教出來。畢竟，大部分的科研工作都是苦哈哈地做出來的。我列舉這些例子，主要是給大家看看：甚麼叫好。如果你覺得這些離你太遠，也不要灰心傷志，因為其它的大部分文章，都不是這類水平。

我甚至覺得，若是在所有有「項目」的學者中，有 2% 的學者能到這種程度，則國家獎勵學術的經費，就沒有白花。要國家獎勵學術的經費沒有白花，談何容易啊！

最後，我用一小段引文來結束這一段。第一段來自《東周列國志》：

姬平乃歸燕都，修理宗廟，志復齊仇。乃卑身厚幣，欲以招徠賢士，謂宰相郭隗曰：『先王之恥，孤日夜在心，若得賢士，可與共圖齊事者，孤願以身事之，唯先生為孤擇其人。』郭隗曰：『古之人君，有以千金使涓人求千里之馬。途遇死馬，旁人皆環而嘆息，侍從官問其故，答曰：『此馬生時，日行千里，今死，是以惜之。』涓人乃以五百金買其骨，囊負而歸。人君大怒曰：『此死骨何用？而費吾多金耶？』侍從官答曰：『所以費五百金者，為千里馬之骨故也。此奇事人將競傳，必曰：『死馬且得重價，況活馬乎？』馬且至矣。』不期年，得千里之馬三匹。今王欲致天下賢士，請以我為馬骨，況賢於我者，誰不求價而至哉？』

國家出錢買科研成果，是不是這樣想的呢？

February 2, 2007

第三部分：一般論文

在功力不足時寫好論文較難，但前面說過不能等一切都預備好了再寫論文。這樣做真可能一輩子都只是做講師助教授。因此，新手上路是必須的。並不是說新人就寫不出好論文，世上大有聰明的人，極年青時就光華燦爛，引人注目。我們當然希望最後大家都有較大的學術貢獻，但目前我們且把身段放軟，將目標別定得那麼高。

如何寫論文是次要的，如何找題目才是真的。現代網路發達，只要蒐尋的方向是大略正確的，年青學者並不難找到相關的學術資訊。

別做學位論文的推廣

畢業論文裡往往有最後一節：未來的工作或今後研究的方向。你當然會很自然地去做原論文的推廣，不論怎樣，這是你還算熟悉的專題，何況你的博導說不定還會給你一點指導。

我的建議：別走這條路；或者，至少你別希望靠這條路多寫幾篇論文。

理由是有一些的：第一，這往往是沒有多大前景的題目。學位論文是用來訓練研究生的研究方法、態度和寫作技巧用的，而訓練的意味大於研究的意味。你的博導知道你大概做得出來，或者至少他自己知道他一定會做得出來。這樣的題目做練習可以，做研究就有一點不夠難。第二，如果你仍然倚靠你的老師，請問論文將來要不要掛上他的名字？如果要掛的話，是你排第一還是他排第一？我其實並不十分反對老師和學生合寫，畢竟老師在學生身上用了不少時間指導，因此在學生的論文上掛一個名字也算應該。但我覺得三篇以內還算可以，超過此數就有一點過了——學生更要考慮升等，研究項目等世俗上的事，聯合而寫的論文，尤其是和自己的老師合寫的，在學術評鑑上是會被打折計算的。道理很簡單：升等的最重要的要求是「該員有沒有獨立研究的能力」。獨立者，不靠別人自己單幹也¹。第三，老師的題目，如果他是這個專題的專家，你多半做不過他（和師兄師姐）。青出於藍談何容易？別忘了這些人都是競爭者。

簡單地說：要長大，先斷奶。

數據和論文

統計學的論文分為兩大類：有好數據的和沒有數據的²。做沒有數據的問題比較簡單，作者基本上只要提出，對這個問題，A 做了甚麼，B 又做了甚麼，但他們都沒有做出結果 C。你只要把 C 做出來了，當然就算是研究有成果。這在 A、B 都是有名氣的教授時，你的理由就更充分了。做這類問題，最要緊的是你手上的數學計算的功夫要好。國內出身的統計學者，在國外好期刊上經常有論文的，幾乎沒有人數學不強的。

只要數學夠強，你可以放心大膽地一再把條件減弱或者計算得更精細。例如某人證明了弱收斂，你就做強收斂，或者進一步去算它的收斂速率，甚至大離差 (large deviation) 的結果 ...。

¹在某些單位，「沒有 single-author paper」是研究能力的重大瑕疵。

²如果你只用某些老掉牙的數據，如 sun spot data, iris data, Nile river data 等，不算是有數據。

下面我舉一個例子，是關於 Kaplan-Meier estimate (KME) 的，前面已大略提過。

KME 的故事

給一組「隨機設限 (random censored)」的數據：

$$(\min\{X_i, C_i\}, \delta_i), \text{ iid } i = 1, 2, \dots, n$$

其中 $X_i \sim F; C_i \sim G, \delta_i = I_{[X_i \leq C_i]}$ 。若 $\delta_i = 1$ ，則我們只看得到 X_i 的值，此外 $\delta_i = 0$ ，我們只看到 C_i 的值。主要的目的是尋找 F 的估計量。

如果所有的 $\delta_i = 1$ ，則我們能看到所有的 X_i 的值。此時，沒有因設限而丟掉的值。那麼

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I_{[X_i \leq x]}$$

是最傳統的估計量。關於 $\hat{F}_n$ 的性質，已經太古典了³，而這由二項分布和 CLT 及相關的大樣本理論就知道。簡單地看，這是以 $\bar{X}$ 的形式出現的統計量，能夠做的數學，大概都已經做完了，如果還剩下甚麼，都難得要命。

但是對於 KME 而言，因為有幾個 $\delta_i = 0$ ，有些 X_i 是看不見的（實務上的說法是我們知道 $X_i > C_i$ ，我們只看到 C_i ）。KME 的出發點反而是風險函數 (hazard function)。若 $X > 0 \sim F$ 是一個不會為負的隨機變數 (例如壽命)，則

$$h(x) =: \frac{f(x)}{1 - F(x)}, \quad x > 0$$

叫做 X 的風險函數。利用微積分可求出

$$S(x) =: 1 - F(x) = e^{-\int_0^x h(u) du}$$

如果將上式中的積分用一段一段的求和來表：

$$h(x) \approx \sum_j h(u_j) \Delta$$

則有

$$S(x) \approx \prod_j e^{-h(u_j) \Delta}$$

是一個由「求連乘積」的方式來表現的。KME 的結構，就是一個連乘積的形式

$$\hat{S}_n(x) = \prod_{t_i < t} \left(\frac{n-i}{n-i+1} \right)^{\delta_i} \quad (1)$$

其中 $t_i = \min\{X_i, C_i\}, i = 1, 2, \dots, n$ 。

公式 (1) 的源遠流長，但之前的工作，如涓涓細流，可是到此之後，氣勢一變。關於對設限數據求 $F(x)$ 的估計的問題，甚至於一般地對於設限數據的統計推論問題，都有了開朗的看法：原來我們得從風險函數入手，而不是從「算術平均數」的角度來看問題！這

³意思是，別在裡面找題目了，因為幾乎都挖光了。

就像人類一直都在用模彷鳥類的角度來設計飛機，一直沒有成功過。直到萊特兄弟的定翼飛機成功了，人類才忽然懂得：不能學鳥⁴。因此航空工程才能突飛猛進一樣。

Kaplan 和 Meier 的文章，好像是忽然將問題的「任督二脈」打通一樣。自然後面就跟了一大堆補強、加註、推廣的文章。那是五零年代的末期，搶做論文的風氣還不太烈。下面我舉的例子，多多少少可以讓大家看一下怎樣去搶（還不算太糟的）題目。

有了 KME 這樣的公式，甚麼時候可以大方地用它？KM 的 1958 論文中的條件是 random censoring，這並不能一定保證，在實用上能不能更寬些？一個較好的結論是：如果「設限」這件事不會影響到未來的壽命，則 KME 的公式是可用的。用數學講，就是

$$P(X \in [t, t + dt] | X \geq t) = P(X \in [t, t + dt] | X \geq t, C \geq t). \quad (2)$$

這樣的條件有明白的直觀意義，比 random censoring 要寬鬆⁵，這是蠻有意思的結果，當然值得發表⁶。

但這樣的結果就沒有問題？記住，「找題目來做」的技巧，有一部分就是挑毛病。要挑毛病，即使別人沒有毛病，也要努力去挑。有人問：這樣的條件 (2) 是不是可以從所給的數據裡檢驗呢？這不是無聊的問題，即使答案是負面的，也有它的意義。答案真是負面的，當然能講清楚這一點，也是可以發表的⁷。這個結果當然重要，因為它，我們只能利用數據的背景來討論是否可以用 KME。因此，取樣前的行政作業規範，就變得重要。

你可以更進一步，挑出 KME 估計得很差的情形來說⁸。這樣的論文，當然需要 KME 已經被公認重要時才有價值。否則，我隨便發明了一個估計量，或者一個根本用不到的估計量，好或不好都不重要，有甚麼好討論！

較容易問的問題是它的數學性質。這裡有一大票論文⁹。最直接的是 normal approximation¹⁰：

$$\sqrt{n}(\hat{S}_n(\cdot) - S(\cdot)) \rightarrow \text{Gaussian process}$$

證明某統計方法數學性質的工作是可以做的。但愈來愈難做，愈來愈難發表，而且學術上的影響愈來愈小。那樣難，除了那幾個不太識人間煙火的象牙塔中人，誰看哪¹¹？當然，如果你真的快手快腳，又有幾下獨門絕招，這遊戲還是可以玩的。

做這類學問，需要效率。你需要眼明手快——眼到、手到、心到。這是因為有極大批學者在幹類似的事¹²，因此只要有一點有氣味的風吹動，這批人（包括他們的學生）就可能撲上來搶做。你若是慢絲條理的，這口飯就有點難吃。

跟別人搶快的另一個條件是題目紅火的程度。一個好專題在開始的時候，尚待開發的部分還多，你只要夠勤快，機會也會有。過了一陣，這個專題中容易做的部分很快地被

⁴另一個原因是當時的工業水平不夠，機械效率太差。

⁵並且，很多其它的設限方法，在這樣的條件下也可以用。

⁶Kalbfleisch and MacKay (1979). *Biometrika*, 87-90.

⁷Tsiatis (1975). *PNAS* 72, 20-22. *PNAS* 指的是美國科學院的期刊，impact factor 是很高的。

⁸Peterson (1976). *PNAS*, 73, 11-13. 抓到名人行為不檢，也是有市場的。

⁹這我就不一一列舉了，找一篇較近的關於 KME 在 *Annals* 上的文章，從它引用的文章裡向後找，總能找到一大群論文。

¹⁰Breslow and Crowley (1972). *Ann. Statist.*, 437-453. 到了現代這種題目已搶不到了，因為作者若是自己不證，多半論文就登不出去。研究工作者的日子是愈來愈難了。

¹¹又一說法是：很多期刊是掌握在這類學者手上的，因此這類論文反而更有機會。

¹²這是支持學術期刊的主流人群，因為要做有真貢獻的工作不易，做看起來有貢獻的且又挑不出毛病要容易多了。數學暗示困難和嚴謹，這至少就是 non-trivial。

一掃而空。也許這個問題已經被充分了解，學術期刊的編者會覺得這個題目不再有趣並且還有更重要的熱門專題要討論。聰明的人就會考慮下車，尋找下一個目標——以上是「業績掛帥」的幹法。對學術不是沒有貢獻，但一般沒有大貢獻。愛因斯坦說，他最不想見的，就是一大堆學者拼命在軟木頭上鑽洞！

當然，即使在問題的晚期，你也可以做出一些能發表的結果。但僅就投入/產出比來看，是比較不合算些。有時你也可以做出相當重要的結果出來：如果 A, B, C... 等作者，在討論別的部分都很清楚，但到了某處你想弄清楚的地方，言詞就東閃西躲沒一句有氣力的話時，你就知道那是一個需要突破的關鍵點：如果你能將它搞清楚，做得出（包括證明做不出），總是會有好處的。

先拿到數據再做問題

首先我得說拿到新數據並不容易。工業上的數據難拿到，因為廠家有業務機密要保護¹³。其他科學家辛苦由實驗而得的數據，更是他們的寶貝¹⁴。而過去已發表在公共領域（public domain）的數據，多半是原始的資訊，都已榨乾。拿來驗證你的新方法可以，但對於原來的科學問題，已貢獻不大。因為別人的功力也不會差，你不見得會分析得比前人要好。

較好的數據是來自你參加別人的研究團隊。這樣，你就在問題的早期就切入，參與數據的規劃蒐集，和專業人士直接對話。何況真實數據，即使是經由統計學家的設計蒐集，也很難不會發生問題：取樣的獨立性不能保證，資料未必完整，套不進現有的模型，殘差的變異過大或者全然不像是常數...。只要你投入，就會發現教本上的技術都用不上，而這就是你發展新工具的機會。

這是一個統計學者的正途：在別的科學研究裡找到研究的生命力。一開始要打入別人的團隊較難，但等到你漸漸有了聲望、名氣和口碑，上門求教者會讓你有做不完的項目。但你要注意到別做簡單的工作，那樣你只是別人的雇用人員。已故的名生物遺傳統計學家李景均¹⁵說：“I don't do chi-square”，意思就是說：要挑嘴，不要甚麼都吃。

最糟的事是：辛苦做出一點結果，投稿時評審問：有沒有 real data？你東翻西找，哪裡可以找到剛好可以套進你的新模型的實數據？這在你做純理論的工作時還好，至少還可以模擬。在你做號稱應用的統計問題時，這就夠你喝一壺了。初入行者最易犯此病，切記切記！

有真實數據的論文，其實是比较有意思的。所謂數據，最好有重要的學術上的或者生活上的背景。例如你如果弄得到中國衛星掃描而得的數據，或者奧運選手的訓練數據，就算你用較淺的統計方法來分析，只要基本正確，英文再清楚些，發表不難。如果是神六數據或者關於青藏鐵路的數據，當然就更別說了。

數據最好不要太標準。如果套進迴歸軟體就有好結果，這樣的論文誰會要？你需要複雜些的、有結構的、可以至少用好幾種近代模型的、最好背後還有一些有科學意味。一開頭，模型簡單些不妨，然後漸漸放寬條件，做較複雜的分析——有一組好數據，可以玩好幾年呢。下面我舉一個自己的例子。

¹³一種是真的機密，競爭對手可能因此而得利；另一種是假的機密，他們其實是在用簡單的手法，但也不想讓外人知道。

¹⁴實驗室間的競爭，因為和項目經費有關，也很厲害。

¹⁵李景均（C. C. Li, 1912-2003），原是北大教授，1949年後因為規定要用唯物的觀點教遺傳學，跑到美國。寫論文固然重要，執著於學術的本質，才是真科學家。前人遺風，心嚮往之。

火工件的故事

火工件指的是有火藥會爆炸的東西。爆炸通常由某物的引發，如通電。電有大小強弱，因此引發與否，是和 stress 有關的。這問題是和計量響應 (quantal response) 是一類的，只是火工件的對象是爆炸，後者是白老鼠餵藥¹⁶，看幾週後腫瘤是否夠大。

對於計量響應，有興趣的部分是分布中間的部分：多少藥量，可以讓 50% 的老鼠產生夠大的腫瘤¹⁷？對火工件，有興趣的部分是分布的尾部：多大的電流，可以讓 99.99% 的火工件都引爆？

我能拿到一組真實數據是因為手上有一個項目。關於 binary response 的數據，至少有一甲子的研究了，書都有好幾本¹⁸。因此在一九九幾，理論上可做的研究不多¹⁹。

有了數據，如何寫論文？應付原來的委託單位 (client) 不難，我們只是把近 60 年的東西補足了，告訴他們到目前為止最好只能做到哪兒，再出一冊中等長度的報告就好了。不論怎樣，原委託單位只是靠一本美軍手冊，而那手冊主要的理論根據是 1948 的一篇論文²⁰。我們先將問題套進一個 Markov chain 中 (看起來就較深了，因為沒有 independence)，然後再發展理論，用自助法 (bootstrap) 來做所要的信賴區間，並為了提高模擬的效率，引入 important sampling 的觀念，搞出一個自己叫 important re-sampling 的法子來做。經過模擬，用傳統大樣本方法所得到的 95% 信賴區間大約只有 90% 左右，而我們的法子真的差不多做到 95%²¹。

這有甚麼了不起？學問上是沒甚麼：論文裡很多部分的技術都是在文獻裡找得到的²²，但我們有真數據²³，真問題，發現老方法的不足，並合成一些較現代的方法 (bootstrap 和 important sampling) 來改進老問題，這些「功勞」就足夠我們覺得可以寫一篇不錯的論文了。這就是先有數據再動手研究的好處。

這只是一組小型數據的故事。但如果你參加了一個 3-5 年的計劃，其中有 5000 份問卷，分三年完成，那你可以拿來玩一下的數據就夠用好一陣子了。

需讀書時要折節

關於科學意義，如果這個科學有較大而長期的潛力，你就該考慮用心補上一些足夠的背景知識。例如生物遺傳是近年來最紅的專題，就有些極好的華人統計學者，自己找生物教本來弄清楚基本生物學——因為這樣你才能面對著生物學家來談他們的問題。更重要的，你較會知道他們的問題裡面，生物學的意義有多少。

氣不可衰

一般論文指的是雖不見得是最好的，但最後還是在最好或中等以的期刊上能發表的論文。因此，請勿小看它——沒有這許多一般論文，也襯托不出那少數幾篇突破性的結

¹⁶藥量就是 stress。

¹⁷這叫 LD-50。

¹⁸國內有劉寶光 (1995):《敏感性數據分析與可靠性評定》，國防工業出版社。這書蠻實在的。

¹⁹Jeff Wu 有幾篇從 DOE 的觀點來做 binary response 的文章。這也是選題的一法：把腳硬套進鞋裡，套得進就是結果。

²⁰Dixon and Mood (1948). *JASA* 43, 109-126. 所以，你該有一點懂，所謂「機密」，有時會是甚麼事吧？

²¹Chao and Fuh (2001). *Statistica Sinica*, 11, 1-21.

²²其實，論文大體都如此，哪能全都是新的？有一兩點新意就不錯了。

²³指首次使用。

果。古人說千軍易得，一將難求。像岳家軍、戚家軍這樣的精兵是可以訓練出來的，但岳飛、戚繼光卻是天生的將才，這種菁英自己會成長，用不著我們去擔心。研究的結果也一樣。好結果難，多半的結果都是一般的。但它們做了一件重要的事：把問題週邊的枝枝節節都清掃乾淨，因此留下未做的部分都是較難啃的部分。

因此，即使你也在做這類「搶食大餅」的工作，也不必輕看自己。因為你畢竟還有些貢獻。何況這類工作並不容易，世上的高手如雲呀。

做研究除了本身有趣之外，就是還得撐著一口氣別放鬆了。做一些大家都做得來的工作並沒有太糟，但你不能一輩子總是這樣幹。

你也許在想，等我升到教授，就改走行政路線，系主任、所長、院長...一路幹上去。話是不錯，的確有人的生涯規劃是這樣的。但是世事難料，別忘了在大學，教授是本職，其它行政業務都是兼任²⁴。如果你到了五十歲，爭取某一個長失敗，前一個長任期已滿，不得已還是得回到本來的系，那時候年青的學者已趕了上來，正是「長江後浪推前浪，前浪死在沙灘上」。這時，若是沒有一些有功底的作品，你在那個單位裡如何走路有風？

在學術單位裡，學問是一輩子的事。

做較難的問題較有成就感

這話當然不必我說就懂。你可以做幾篇或甚至幾十篇搶快型的文章，但心中最好還是有一兩個較難的、希望能突破關鍵的問題擺著。在某一個專題久了之後，週邊的問題都搞清楚了，偶然來一個突破還是可能的。

我一般不鼓勵年青學者做大猜測 (conjecture)²⁵ 或有名的未解問題 (open problem)。小教授提出的 open problem，學術上的份量也許不夠，做出來功勞也不大。大教授提出來的，往往真正困難。他做不出²⁶ 多半你也做不出。下面的問題是 open 的，感謝林國棟教授給我這個題目。

給一個小題目

令 $F(x) = (1 + e^{-x})^{-1}$, $x \in R$ 。這是所謂的 logistic 分布的 cdf。取 $X_1, X_2, X_3 \sim \text{iid } F$ 。再令

$$X_{(1)} < X_{(2)} < X_{(3)}$$

為它們的次序統計量 (order statistics)。取

$$U = X_{(2)}, V = \frac{1}{2}(X_{(1)} + X_{(3)})$$

則可以證明 U, V 有相同的分布。但這個命題的逆定理是否成立？i.e., 若設 $X_1, X_2, X_3 \sim \text{iid } F$ ，取 U, V 如前，如果 U, V 的分布相同，問 F 是否恰好是 logistic？

這題目我不建議你去做，道理很簡單：林教授在這類問題上至少有二十年功力，他若是還做不出，肯定是難的。當然，你若是能嚴謹證明，發表應該是沒有問題的。

新意

²⁴起碼台灣是這樣。

²⁵好在統計學裡沒有甚麼大猜測可以讓人一夜成名。

²⁶多半他的門人弟子也做不出。

競爭的目的是汰弱留強或者變弱為強。光是要求 SCI/SSCI 期刊, 已相當能夠汰弱了。這些期刊對於國內學者而言, 大都沒有甚麼人情可講, 如果論文沒有太多的新意, 英文講述能力再差一點, 這對於國內上千個大學裡的學者教授, 已經是較難的障礙。這裡面, 最難的部分是新意。新意有大有小。好論文有大新意, 前已述之。一般論文多半只有一點點的新意。

你怎麼知道你所做的東西是新的²⁷? 主要的是: 你必須知道你現在做的研究專題的現況, 才能知道你想做的方向、你的點子 (idea), 是不是已經被別人做過了, 是不是做得夠好。

這在大家不怎麼重視論文發表時²⁸, 情況還好。但現在大家一窩蜂要求發表論文, 競爭的對手太多。別人不見得笨, 壓力不見得小, 研究資源和環境不見得差, 你看得出想得到的點子, 別人也多半能看得出想得到。如果問題的難度不大, 除非你的速度比別人還快, 你的機會也不見得有多高。國人所佔的便宜是數學訓練較強。這在以前是大便宜, 但現在太多的華人學者在搞統計 (和相關學門) 的研究了, 這些人也不見得人人在前五名的期刊上發表, 因此競爭下來, 國內學者也未必在「數學較強」上佔優勢。

所謂「新意」, 因為人人都在要求, 所以一般對於新意的要求, 其實也沒有那麼嚴: ——你只要能清楚說明你的新意在哪裡就是。重點是你得說得明白, 而且在論文的前兩頁就說明白。

緒論的重要性

這裡說到了論文中最重要的部分: 摘要 (abstract) 和概論 (introduction)。摘要因為短只要說重點就好²⁹, 問題不大。概論則不然, 這是最難寫的。

當學者埋頭做問題的時候, 往往是當他循著某些思路一路向前, 尤其是在有一點靈感的時候。如果你做到某一個深度, 做研究基本上還是有趣的事, 有時顧不得去找是否別人有類似或相反的想法。但這在論文的概論裡, 你就得在兩三頁的範圍裡講清楚這個題目從第一篇論文起到目前為止的概況。概況不是目錄, 你得說出 (用較乾淨的英文) 誰做了甚麼, 結果有沒有影響, 誰貢獻最大, 誰又跟在後面但也有功勞, 這些與你目前論文的關係——這一切只是要表現你是如何的專業。要用明白的話說明你的貢獻在何處, 但又不能用自己誇自己的句子。這一方面是禮貌, 一方面是規矩: 論文要用大心機寫, 但得表現出平常心的樣子。平常心, 指的是客觀和不帶感情。

這裡面學問可大著哪, 一時也說不完。當你漸漸有名而在較好的期刊上投稿時, 別忘了你面臨的評審者都是這一行的專家, 並且, 更重要地, 他們都是你的競爭對手。「行家一伸手, 便知有沒有」, 你拉出來的架勢, 已差不多決定勝敗。因為大概由這裡已可看出你是一個青澀的菜鳥還是一個戰鬥經驗豐富不可輕侮的高手。何況這裡面還包含著學門和學派間的恩怨³⁰, 因此有些文章你是要捧的, 有些是你是要罵的。捧的時候你要維持做一個學者的尊嚴, 要恰到好處而不帶諛詞³¹, 罵的時候要尖酸刻薄不留餘地地突顯別人的缺點但還得顯出溫柔敦厚純學術本位的樣子。你以為只要結果好就沒有問題, 是

²⁷一個方法是先查一下 Kotz and Johnson 的 *Encyclopedia of Statistical Science*, 計九冊, 是一個蠻中肯的開始點。可找到很多專題過去的關鍵論文。再去反查 *Citation Index*, 便可回溯到近年。這兩種資料都貴, 要靠一個好圖書館。至於現在別的競爭者的工作, 那就全靠你自己的本事了。

²⁸如五十、六十年代。

²⁹據說人大要求學位論文的摘要至少三千字, 是有點多。這樣, 誰還讀論文的本文哪?

³⁰你以為拒絕過某人或某人學生的文章, 別人真的不知道? 學術圈子其實很小, 有些人真的不怎麼上道。

³¹有一次某文章提到我的某論文, 說我做了 A 又做了 B。但我確實沒有做 B, 反而因此拒絕了那篇文章: 他沒看懂我嘛!

的，真正的好就大概就沒有問題。一般的好，還得看江湖上的朋友是不是肯高抬貴手。

同理，別人文章的緒論，也得好好讀。第一次不妨馬虎些，讀完全文以後，如果你覺得這文章還有點意思，再回頭讀一次該文的緒論。問自己三件事：第一，這文章的主要賣點在哪兒？第二，作者怎麼想出來的？第三，現在你已經「懂了」，請問若是由你來寫，這緒論該如何寫？

這和研究學問的關係不大，但和你的論文接受率有關。懂得化妝的女子，找到愛情的機會大些。投稿相當於「求售」，即使是和氏璧，在未打磨包裝前都一再碰壁。文章不注意潤色就投出去，等於穿得一身邋遢去做面試會談。——便縱有千種風情，更待與何人說？一篇論文的緒論，和一個人的面孔一樣：別人第一眼就要看的，而印象好不好對下一步決定的取向，可是有極大影響的。

要有所影響 (make an impact)

最好的論文開創新域。大如相對論，對整個的科學投下重磅炸彈。小如前述的 information theory，也開拓一個特殊的專題。其次的論文，在突破一個重要的關卡之後，前面是一馬平川。例如 Ito's lemma，為隨機積分鋪下定石。再次是當某類問題已日漸成熟，經過深思熟慮的梳理之後的集大成之作。

能做這類論文的學者，根本用不著我們去擔心。但即使是一般論文，也該極力地求新求變，希望能對原來的學門產生影響。我們且不談真正的科學貢獻³²，只是為了世俗的目的：為了能在較好的期刊上發表，得到較多的引用 (citation)，得到同儕的認可，得到國家的獎助，得到學術上的榮譽等等，你就得出盡絕招，求新求表現求有所影響。因為「是不是有影響」才是我們對學術研究是否有意義的基本考量。

判斷一篇文章有沒有新意，可從幾個簡單的角度來看³³：是否有新的資料，新的觀點，新的切入點，新的分析方法，新的發現。

判斷一篇論文是否有賣點，也是看它是否能有所影響。因此作者要用心去強調他的論文確有影響。當然，有沒有影響是在論文登出來之後，別人才會去讀，而且讀了後會去用才會漸漸地為學界所知。因此目前誰也不能真的確認，也因此你就該誠實³⁴地極力鼓吹³⁵。

在論文裡不要謙虛。不要說自己的結果不重要，儘管你心裡也許會覺得如此³⁶。外國人常明白地告訴學生：“Be positive !”，多講你的成功，別說你的失敗。

懷疑論

做研究時的基本態度是誰也不要相信，甚麼也別相信³⁷：例如一般的論文中號稱的好結果，別人給出的 conjecture，論文裡的定理和應用，你的系主任，院長甚至博導所告訴

³² 一般的論文，本來貢獻就不可能大。貢獻大的都是好論文，這是定義。

³³ 以下幾點由管中閔教授的一篇講稿中引用。

³⁴ 不要講你自己都不信的話，專家還是很多，還是要回到學術面。因此，沒有狼，不可以說狼來了；但若有狼，可以說大惡狼來了。

³⁵ 中山先生革命時，若是很謙虛地說滿清政府雖然腐敗但仍擁有極大的資源，參加革命多半會死，誰會陪他去玩命？但若是將建國大綱，五權憲法這些搬出來，再說滿清是唯一的阻礙，且已千瘡百孔，則前景就完全不同。七十二烈士有幾人讀過三民主義的全文？他們多是只聽到了孫先生的 introduction!

³⁶ 這在你原來有一個大目標，但只做出來一個次要目標時，心中多半會覺得如此。

³⁷ 這一段是利用林共進教授的講稿改寫的。

你的事。因為他們可能會錯，會不周延，會省掉某些條件，會條件太強，應用範圍不廣。

爲甚麼要懷疑？因為如果這些人都是對的，都已做到最好，那麼你除了趕快換題目之外，還能做甚麼？因此你必須用有色的眼光來讀別人的論文，用懷疑的態度來看科學的命題³⁸，用別人沒有問過的問題來拓展自己的深度。題目要自己挑，因為別人給的，可能真是殘羹剩飯。

但天底下甚麼題目都能做，甚麼都不能做。明明是死棋，你也可能走活；明明是活棋，你也可能走進死胡同。信心和研究能力都是靠跌跌撞撞培養出來的。

再說懷疑論

這等於是上一段話的 corollary。因為別人論文的好處別人自己會說，因此他們不提的部分——也許是忘了，也許是沒有想到，也許是故意³⁹——才是你該注意的。Kaplan 和 Meier 沒有說在隨機設限之外，KME 是否可用，這就成就了 Kalbfleish 和 MacKay 的文章；而 Kalbfleish 和 MacKay 也沒有說出他們的條件，是否可由 KME 的數據來檢驗，這就成就了 Tsiatis 的文章。

這些文章的題目都由懷疑而來，這類的結果，不是挑錯⁴⁰，而是補強、拓展和更清楚的說明。因此即使文章交給原作者去審，多半也不會引起反彈。

February 2, 2007

³⁸例如在某演講裡聽到某問題可以做或者是 open 時，就要多考慮：多半是雞肋。

³⁹包括他自己還留一手預備寫下一篇論文。

⁴⁰理論上，如果你說某人的結果是錯的，他應該「聞過則喜」才對。可我就聽過這類文章被一票人圍堵的故事。江湖中，固然有俠士，也有拉幫結派的黑道。

第四部分：動手作文

沒有研究結果就沒有論文。有了之後，寫論文只是將你的研究結果，清楚地告訴別人而已。對學術論文而言，你的讀者是學有專精的人士，你的結果是實實在在的科學結論，因此你不是寫詩，不是寫散文，甚至不是一般的論文¹。學術論文要說得清楚明白，沒有不相干的話，基本上要以客觀的立場敘事，而且遣詞用字以不帶感情為宜²。

邏輯次序

假如你已得到 A、B、C 三個結果。你在做研究時，因為你對於問題背景的熟悉，思想方式可以跳來跳去。但是讀者雖是專業人士，卻並不是你。因此你得將結果的邏輯順序弄清楚。

簡單地可以這樣想：如果你要將結果講給學生聽，你該用甚麼次序？是不是先弄懂 B，才能弄得懂 A？在 C 中所用到的幾個名詞，在甚麼地方引入最好？「你的學生」是還不錯的模擬對象。因為他們有點懂，又不真懂。至於讀者的程度，可以假設他們有一個計量方面的博士學位。

先粗具規模

寫論文也是作文，但目的是講故事：將你的研究結果講給別人聽。因此，成果第一，故事第二。先將成果依邏輯次序列出來，然後用文章將它們連起來。將想講的其他各要點一一列出，試著替它們安排適當的出場點。前幾次的稿子可以疏漏，但你得一遍一遍地潤色，到最後打磨光亮為止。

這樣的方法，適合於你自己用電腦軟體來打自己的論文。因為在軟體上修改和打印都十分容易，所以你才能一次又一次反覆地修改到滿意為止。這和幾十年前，你要把手稿交給打字員完全不一樣。因為那樣一來一往，至少一個星期。我們要盡量發揮近代科技的優點。世上有些學者的本事大，他們第一遍的手稿就和最後的定稿差不多。如果你有這樣的本事當然最好；如果沒有，不妨學我。自從開始用軟體寫作以來，一篇論文 update 十次，是常有的事³。

初具規模就去動手拉開文章的架勢的好處是：它讓你「下決心去寫」這個步驟，變得容易⁴。

萬事起頭難。若要不難，胡亂起一個比總不開始要好。

從第二節開始

寫論文要從第二節開始寫，用心讀論文也是從第二節開始。何以故？因為一般的論文，第一節的緒論只是誇說自己的結果多麼好，別人是如何有缺點而已。真功夫在後面

¹例如報紙上的社論，或者《古文觀止》上的那種古典論文。

²有時，像「感謝某人的討論」、「慶祝某人七十生日」之類的句子是可以加上的，但不在正文。在正文裡要一本正經討論學問。

³當然，A4 的紙就用了不少。可是這是你項目經費裡最該花的錢，並且不能算貴。

⁴我做學生時，老師的教導是：沒有好結果就別寫——寫爛文章是丟人的事，有尊嚴的學者是不做的。但現在政府如此推動發表論文，大家有甚麼辦法？我現在只能主張大教授別甚麼論文都寫，總要有點品味罷？

(有時甚至在某一個放在附錄的 lemma 裡)。緒論需要用心化裝鋪排,但化裝鋪排也會掩蓋缺點⁵——要真懂,還得自己去用心讀。

最後才去寫緒論。緒論 (introduction) 指的是「介紹」——將你的內容介紹給讀者認識。因此先有內容,才能介紹。

基本格式

論文的基本格式,大體相同。雖然各期刊都會有一些要求,投稿之前,去找一本近期的目標期刊,讀一下它的「Instruction to Authors」,就差不多知道了。你其實還得去讀一下該刊的「編輯方針 (scope)」,他們多半會說明他們喜歡哪一類的文章。

論文多半以節 (section) 來分,而不是以章 (chapter) 來分。章比較長,在專書裡常用;而節較短,因為一篇論文,一般只針對一個專題。

節之下可再分小節 (subsection)。有些人還要分小小節 (sub-subsection),我覺得是太細了⁶。好一些的論文,在某節節尾,常會說下一節會討論甚麼——語氣是隨著本節的內容來說的,給人的感覺是:每一節都有 motivation,一個忌諱是:某節的標題後面,沒有說明就加上小節的標題,如

4. Estimation Procedure

4.1 Estimation equation

下面的寫法較有樣子:

4. Estimation Procedure

At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here. At least one paragraph of text here.

4.1 Estimation equation

不少人不曉得這事,雖然現在已不那樣講究,但是一連兩個標題連在一起,當中甚麼內容也沒有是蠻難看的。另外,節或小節的內容都不要太短,因為印出來的版面也不好看——好期刊是非常重視他們的版面的,因為學術論文的目標是「藏諸名山」,如果不够講究,就表示他們自己都不在乎自己的內容。而學者們肯去投稿,除了升等之類的壓力外,另一個原因則是「好期刊是要被好圖書館收藏的」——這和在網站上隨便貼一下完全不同。

至於圖和表在文中該放在何處,才較好看,你不必去擔心,因為這是排版者的專業。你不妨將圖和表都放在文章的最後,只要在正文中用一行字說

(Insert Table 3 about here)

⁵記住,誰也別信。

⁶如果有 3.4.4 節,多煩哪。

就可以了。

文章是給人讀的，因此要有次序。學術論文也是要給人讀的，但讀分兩種：粗讀和細讀。

你的論文摘要，要能讓讀者有興趣進一步去讀你的緒論。你的緒論，應該能讓讀者去粗讀你的正文。至於細讀，只是針對那些可能用到你的論文的人和你的競爭對手。刊登論文的主要目的是留下紀錄 (documentation)，是先驗證過 (所以要學術評審) 再給人查的 (所以圖書館才收藏)。因此別太希望有人會用心細讀你的論文，但你得希望有人會粗讀它，抓住你的結論。因為極少人會從頭到尾將每一個字都讀到⁷。

一個方法是以論文的主體作為讓人粗讀用，而將較細膩的細節，放到附錄。如果期刊是有聲望的，多半的讀者會覺得附錄中的東面是評審查證過的，因此，除非必要，他們也不會細看。

同理，不要把你的每一招每一式都寫出來。我有一篇文章⁸，其中的一個公式的手稿就有 100 頁，而我算了四遍。但寫入文章，不過九行。我只是淡淡地說：

The computation, although tedious, can in fact be carried out. (以下為推導的概略方法及最後公式)

沒有這個公式，這篇文章大概不會接受。但若我寫上 20 頁，恐怕更難被接受。

平等的精神

在科學之前，人人平等。這表現在論文裡，就是一般在我們提人名時，不提他們的職位、榮銜或者尊稱。我們說 Smith (1999)，指的是此人在 1999 年所發表的一篇論文或者專書。但我們不提此人是大教授、小教授、新出爐博士、某科學院的院士、還是某國國王所封的爵士。因此，文章中別用 Dr. Taguchi, Professor Ito, or Sir Fisher⁹ —— 要讓他們的文著作自己說話。即使是自己的老師，也是一樣。文章裡不作興拉關係，說誰是我的師叔，誰又和我在某會議中相談甚歡云云¹⁰。雖然，圈子裡對大家的關係都略知一二 —— 你以為大家去參加國際會議只是去聽別人演講？

這並不是說我們不承認好文章的權威¹¹。但別人的 credit 要靠他們的論文來展示，不是靠頭銜。「平等」的另一個意思是：至少在表面上，你的文章是和大家站在同一標準來評量的，不因你是新手，就該「讓幾個子」¹²；也不因為你聲名赫赫，就不敢懷疑你的結論¹³。

嚴謹的態度

在所有的嚴謹中，最要緊的當然是學術推理的嚴謹。當它有漏洞而被評審人員抓到

⁷現在的論文真是太多了，誰有那個美國時間哪？

⁸Chao (1987). *Biometrika* 74, 426-429.

⁹這些人拿到的榮譽已經夠多，用不著我們來湊熱鬧。

¹⁰但文章中可以感謝某人提供建設性的建議之類，這類事該放在「感謝」這一小段，不可放在正文中。且不宜感謝期刊的總編和副編，因為有拍馬的嫌疑。

¹¹我們對好結果還是有敬意的，文章中不妨適當表示。但別過分。

¹²有些期刊對博士論文較鬆，這是他們的政策：不要嚇壞了年青人。

¹³有些期刊採用「寄出評審時將作者姓名塗去」的辦法，以免有名的學者有較高的機會通過。但這其實不易瞞過內行人。

時, 除非你能夠補救¹⁴, 否則你只好摸摸鼻子走路。

其它的嚴謹就表現在許多小地方了。我們列舉一些。

英文要正確

英文的文法別錯¹⁵。該加上 s 的地方, 別忘了加上。單字別拚錯了¹⁶, 因此自己細讀幾遍是必要的, 千萬不要以為評審人會幫你一一挑出來, 他又不是你的博導¹⁷! 拚法對的字, 未必是正確的字。實在不行, 找個英文好的老外 (包括付費) 將它順一順。

國人的英文多靠上課所學。學得好的人, 寫出來可以讓老外能懂。但英文的精微細緻處, 沒有十幾年功力是不行的。比如說, 你知不知道, can not 和 cannot, 哪一個對? Disqualified 否 unqualified 有甚麼不同?

另外的要求是一致性。例如我在文章中說, 1998 年, A 做了 A1; 2002 年, B 做了 B1。這兩個句子, 動詞的部分, 可以用過去式 (98 和 02 都已過去); 也可以用現在式 (學術上的結果可以以真理視之)。這兩種寫法, 現在在論文中都可以用。但是, 在同一論文裡, 不要在某一句裡用現在式, 又在另一句裡用過去式¹⁸。這是文法上的不嚴謹。

關於英文, 寫到此處。我推薦一本極好的小書。當年我極為受惠, 而且至今還是經典。網路上一查便得。

William Strunk, Jr. *The Elements of Style*.

我只是將它的目錄列在下面:

Elementary rules of usage

Form the possessive singular of nouns with 's

In a series of three or more terms with a single conjunction, use a comma after each term except the last

Enclose parenthetical expressions between commas

Place a comma before and or but introducing an independent clause

Do not join independent clauses by a comma

Do not break sentences in two

A participial phrase at the beginning of a sentence must refer to the grammatical subject

Divide words at line-ends, in accordance with their formation and pronunciation

Elementary principles of composition

Make the paragraph the unit of composition: one paragraph to each topic

As a rule, begin each paragraph with a topic sentence; end it in conformity with the

¹⁴如重新證明, 換一個較嚴的條件, 或者證明評審錯了。

¹⁵我們假設你是用英文作文 —— 你要 SCI/SSCI 是嗎? 雖然, 中文的論文, 對文法的要求是類似的, 你只是沒有問題而已。

¹⁶現在的軟體都有 spell check, 一定要用! 雖然用過之後還會有錯, 但要好得多。

¹⁷有一個故事是這樣的: 某生在他的論文某處加上「若老師讀到此處, 請找我, 有 XO 一瓶為謝, 以感謝老師的用心細讀。」—— 據說 XO 還在。

¹⁸即使兩句是隔了 20 頁也一樣。這個毛病在作者參考了幾篇別的論文後再寫緒論時常犯。你以為洋人都是對的。不錯, 他們分別都是對的, 但你合起來寫就錯了。

beginning
 Use the active voice
 Put statements in positive form
 Omit needless words
 Avoid a succession of loose sentences
 Express co-ordinate ideas in similar form
 Keep related words together
 In summaries, keep to one tense
 Place the emphatic words of a sentence at the end

A few matters of form

Words and expressions commonly misused

Words commonly misspelled

其它細節

細節是常忘但忘了的結果就是 sloppy 的東西。我列舉一些於後。

一字一義

在全文中，一個名詞指一件事。不要在第二節裡，你的意思是 A，到了第四節，你又指 B。爲了英文顯得活一點，你在文章中可以用一些同義字，但不要在有技術性的字彙上搞這一套。

同理，一個符號只代表一件事。如果你用 $f(x)$ 來表示 X 的 pdf，那麼後面的文章裡，你就不能用同樣的 $f(x)$ 來表 Y 的 pdf，除非 X, Y 是同分布的。理論上，全文的符號都要一致。這毛病在你組合不同時期的手稿時會犯，自己往往都不會覺得。

新詞要交待明白

所有較難一點的詞句，在第一次出現時都要給定義¹⁹。不一定是非常正式的定義，但要能說明白。

相互引證的事情要一一對應

例如正文中口提到 Wang (2002)，那麼在最後的 References 裡，就該將該文（或該書）列入。反之，若後面的 References 中有 Wang (2002)，則正文中某處需提到以文章或書——兩邊誰多一個誰少一個都是不嚴謹。這原則也用在同一文章中的圖和表上。若有圖三（或表四），則正文中需有某處明示「見圖三（或表四）」。

有些較好的期刊甚至要求你只能引用有第幾卷、第幾期的論文或專書²⁰——某某單位的技術報告的不行的，因為它們往往沒有經過學術評審，份量不夠（也許竟是錯的²¹）。至於網站上的資料，現在尙無定論。我自己的期刊，主張用一個 footnote 來說明出處就

¹⁹請自行判斷何謂「較難」。例如 $e, \pi, E(X), Var(X)$ 可以不說明便還自使用，但 $\phi(N(0,1))$ 的 pdf 就可以考慮說明一下。這雖然已十分通用，但尚未 100% 被認可。

²⁰而且最好不要是不見經傳的小書店的專書。

²¹這就是前面所說的「誰也別相信」的另一種表現。

好,暗示這類的引證,份量要差些。並且,網站有時會消失不見,較沒有「立此存證」的意思。

至於文章中用到甚麼電腦或軟體,以前是有人會提的,但現在已不重要。因為現在幾乎人人都電腦和軟體,就傾向於不提了。這些都已是必需而不希奇的工具了,和原子筆、A4 的紙一樣。誰在乎你用甚麼牌子的原子筆哪?

事事來歷明確

論文用字,如老杜的詩:無一字無來歷。自己發明的,就要明白說出是自己發明的²²,若不是你發明的,你得告訴別人是誰做出來的——別打馬虎眼。這裡若是交待得不明白,將來說不定會有人來檢舉「抄襲別人的研究成果」——這在學術界是大忌,那時說不定教授的資格都被吊銷。

來歷明白的更進一步的意思是:你需要適當地給 credit——一般是引用最早的有突破性的、關鍵性的論文。這是你表現功力的時候:有些論文是有錯的,有時,這些錯內行人是知道的。若是你不查或根本不知道錯而引用,遇到內行的評審,效果適得其反。

在你說明甚麼還沒有做出來的時候,更該用心。文獻蒐尋 (literature search) 本是 research 的一部分,甚至是較重要的一部分。這是初入行者的通病,尤其在他們附近沒有一個好圖書館時,更是吃虧。

不可跳號

論文中的公式、圖、表都有編號。通常各期刊都有自己的規定,參考一下就好。你的編號系統也許和該期刊不同,這一開始沒有關係,多半在論文要排版時,講究的期刊會要求你改(或他們幫你改)。這些不會影響到你的論文接受與否。

但是跳號就會。例如你原有表一,表二,表三,後來你自己刪掉表一,那麼,刪掉後原來的表二,就該晉升為表一,原來的表三,就該變為表二。如果你不改正,文章中只看到表二表三,但不見表一,這樣的跳號,明顯地表示你不認真用心。

拉丁文

為了表示有學問,有些拉丁縮寫是可以用的。例如

i.e., viz., ad hoc, et al., per se, ibid., e.g., ...

但是,別用錯了(尤其是標點,甚至要用斜體字)。

不要隨便就出來一個縮寫

其它的縮寫,應在第一次出現時給出全名。例如「maximum likelihood (ML)」,「generalized linear model (GLIM)」。

總之,英文並非你的母語,英文不好是不能避免的問題。下面的法子有所幫助:

²²而且還得大聲說,生怕評審和讀者看不到,漏了你的功勞。

- 寫不了長句子，就用短而文法不錯的²³。
- 不會玩花樣就別搞花樣。
- 文字上既然贏不了人家，就要以內容取勝。搞清楚你的賣點在哪兒，別人的缺點在哪兒。科學畢竟不是文字遊戲，雖然，文字好還是佔便宜。
- 實在不行，先用中文寫出來再翻譯成英文再改。
- 對於技術性的論文，其實需要用的句子不多。讀別人的論文時，不妨將別人的句型抄下來。若有 10-20 條樣版句型，也就差不多了——畢竟我們不是寫小說、散文或者劇本！

強調緒論

我們一再強調，一篇論文中緒論最重要。從編輯和評審的角度來看問題²⁴：第一印象大體上就決定了他是用正面的角度還是用負面的角度來處理你的文章。他是正面地幫助你呢，還是負面地快快幹掉你的文章？

我通常在兩個小時以內就有了看法——不論我懂還是不懂論文的內容。當我是正面的時候，我才會用心地設法了解內容，這篇文章就有了機會。一半以上的時間我是負面的。如果小錯誤一抓就一把，你會覺得這文章會好嗎？我覺得我還算是夠負責的讀稿人。當我挑到第十個錯（大和小）時，我就會停住了——這個作者已告訴我他不用心。光是 sloppy 這一個缺點，就夠拒絕它了。

最後建議

寫稿及投稿前應去跑一跑學校的專業圖書館。親自去瞧一瞧你的目標期刊是甚麼樣子的²⁵，別人的文章是怎樣的？副編都是甚麼人²⁶？這樣，你比較會知道你的文章是否適合這個期刊。

至於它們是不是 SCI/SSCI，上網一查就知。大體來說，有名氣有歷史的學會和大學所支持的期刊，多半是有水準的。現在有很多由某書店支持的期刊，雖也有好的，一般就較弱。

有些期刊上，有時也有由編委寫出的告訴大家的應注意事項²⁷。我再另外附兩個 power point 的講稿，分別來自林共進，管中閔教授。講的都是關於論文發表的事。

February 2, 2007

²³ 其實最好用長短不一的句法，才不單調。

²⁴ 別忘了這些人都是較有名的學者，他們的一個共同特點就是沒有時間。

²⁵ 否則你的文章是寄給何人，寄到何處？

²⁶ 主要是去查一下這些人最近的著作是否和你的文章有關。

²⁷ 例如 M. Schminke (2004). Raising the bamboo curtain. *Academy of Management Journal* 47, 310-314. 就有參考價值（感謝劉長萱教授告訴我）。

第五部分：且先讀幾個緒論

在這一部分，我們在較好的統計期刊上挑了幾篇論文，將它們的摘要和緒論抄下來和大家討論一下。我在挑選時沒有先讀過論文，我只是跑到圖書館，將架上最近的一期的第一篇論文挑出來。期刊選了四種：*Annals*, *Biometrika*, *JASA*, *JRSSB*。因此不能算是隨機挑選，但也不能算是 100% 的立意挑選。這四種期刊都可算是「第一排 (first tier)」的國際統計期刊。意思是，如果你有兩篇文章在這類期刊上被接受，拿下一個研究計畫甚至升等就有點把握了。習慣上，主編往往會將他覺得較好的論文，排在每期的第一篇。

回應前面我們對於「緒論」的強調，我們試著來讀它們的摘要和緒論。這些文章討論的專題我們不見得懂。但是，就是因為我們不懂，才值得細讀：看看這些寫得好的文章的緒論，有沒有讓我們覺得「後面的文章有內容」？

Green 和 Maridic 的文章

這一篇登在 *Biometrika* (2006) **93**, 235-254. 這個期刊在 1901 年創刊，已有 105 年，是一個有悠久歷史傳承的學術期刊。標題是：“Bayesian alignment using hierarchical models, applications in protein bioinformatics”。

這裡用了幾個名詞。Bayesian 至少大家都聽過。指的是將未知參數當作 random variable 而給一個 prior 的一套想法和做法。Alignment 意指對齊，如士兵的排成直線。Hierarchical model 則是某一類的統計模型，這裡從字面上看不清楚，雖然可以想像，反正是有點複雜的模型就是了。最後說到了應用，用了“protein (蛋白質)”，“bioinformatics (生動資訊)”而個非常當紅的名詞。因此，光是標題，就十分有科學味了。

摘要是這樣的，其上的註，是我加上的。

An important problem in shape analysis is to match configurations of points in space after filtering out some geometrical transformation. (註一) In this paper we introduce hierarchical models for such tasks in which the points in the configurations are either unlabelled or have at most a partial labelling constraining the matching and in which some points may only appear in one of the configurations. (註二) We derive procedures for simultaneous inference about the matching and the transformation, using a Bayesian approach. (註三) Our hierarchical model is based on a Poisson process for hidden true point locations; this leads to considerable mathematical simplification and efficiency of implementation of EM and Markov chain Monte Carlo algorithms. We find a novel use for classical distributions from directional statistics in a conditionally conjugate specification for the case where the geometrical transformation includes an unknown rotation. (註五) Throughout we focus on the case of affine or rigid motion transformations. (註六) Under a broad parametric family of loss functions an optimal Bayesian point estimate of the matching matrix can be constructed that depends only on a single parameter of the family. (註七) Our methods are illustrated by two applications from bioinformatics. The first problem is of matching protein gels in two dimensions and the second consists of aligning active sites of proteins in three dimensions. In the latter case we also use information related to the grouping of the amino acid, as an example of a more general capability of our methodology to include partial labelling information. We discuss some open problems

and suggest directions for future work. (註八)

註一：在這裡提出 shape analysis 的問題，當然大家不會懂。因為畢竟這不是如 OLS 等耳熟能詳的東西，因此要說明一下。重要的問題是 “to match configurations of points in space after filtering out some geometrical transformation” 這不一定是最重要的問題，但他們敢說 important，總有一些根據。此外，在技術面上來說，“transformation” 一詞也提出來了。

註二：此處立刻跳入問題。本文是用 hierarchical models 來架構工作的，而數據中還有標籤 (labels) 的問題，還有兩組要對比的點，數目可能不相同的問題。這說明了問題有難度。

註三：這裡面，何謂 “inference about the matching”，何謂 “inference about the transformation” 都沒有提。但這是只是摘要，是被允許的。

註四：在這裡提出模型，其中 “hidden location”，當然有一點像 latent 的味道，因此用到 EM 就合理了，至於 MCMC，則是因為 Bayesian 的原故。他們更指出，因為模型取得巧，才能有簡化。

註五：再於此處提出他們還用到了 directional data 中的本事，而這又關照了前述的 transformation 有一種是轉軸 (rotation)。

註六：指出只做兩種 transformations，至於 affine 和 rigid motion，一個需要你回到高等幾何，另一個卻是物理。暫時需要想一下，其實是不那麼重要的。但用了較專業的名詞，就會顯得高級一點。

註七：這裡告訴我們，matching matrix (當然，後面的本文一定要定義) 是本文要估計的對象。

註八：最後歸到兩組應用。都是有科學意味的數據，這又關照了標題中的 bioinformatic 一詞。

整個摘要告訴我幾件事：第一，問題是空間中兩組點是否可以相切合？因為這些點指的是蛋白質和胺基酸，所以這是一個當紅的生物資訊問題，這就有了「吸引目光」的效果。然後，作者提出方法，用 Poisson process 來布空間的點，雖然太普通但尚可接受，因為我們也不知道還有甚麼好用的 process。但他們提出了 hidden location，就把問題的架構弄得較難一些了，但他們還能夠簡化，就暗示著「有內容」。第三，他們還從空間統計裡借用工具，並說明他們最得到的是 “optimal Bayes point estimate of the matching matrix”。最後，他們還說：他們用的是真實數據。

摘要給人的初步印象是「有不少東西」。但我們應該看他們沒有說的。例如 “optimal Bayes point estimate of the matching matrix” 是不是科學上有意思的問題要點？或者只是統計學家能做出的東西？從這裡，我們看不出來，matching 做得是好是壞，與是否能最優地做估計是同一件事還是根本不相關？注意到兩個作者都不是生物學家，這是引起我一點懷疑的地方。有兩組數據給兩個統計學家玩，總可以搞些看起來有料的東西出來。估計 matching matrix 和 matching 是同一件事嗎？我怎麼好像覺得作者沒有交待？

現在且看一下緒論的部分。

Various new challenging problems in shape matching have been appearing from different scientific areas including bioinformatics and image analysis. In a class of problems in shape analysis one assumes that the points in two or more configurations are labelled and these configurations are to be matched after filtering out some transformation. Usually the transformation is a rigid transformation or similarity transformation. Several new problems are appearing in which either the points of configuration are not labelled or the labelling is ambiguous, and in which some points do not appear in each of the configurations. An example of ambiguous labelling arises in understanding the secondary structure of proteins, where we are given not only the three-dimensional molecular configuration but also the types of amino acid at each point. A generic problem is to match two such configurations, where the matching has to be invariant under some transformation group. Descriptions of such problems can be found in the review article by Mardia et al. (2003).

We now describe two datasets related to protein structure. One is of two-dimensional gel data, where each point is a protein itself and the transformation group is affine. In this case we have a partial matching, identified already by experts, that we can use to assess our procedures. In the second example we have a three-dimensional configuration of two active sites of two proteins which also has additional chemical information. Here the underlying transformation to be filtered out is rigid motion. In this problem, one of the main aims is to take a query active site and find matches to a given database, in some ranking order. The matches will give some idea of the functions of the unknown proteins, leading to the design of new enzymes, for example.

There are other related examples from image analysis such as the matching of buildings when one has multiple two-dimensional views of three-dimensional objects; see for example Cross & Hancock (1998). The problem here requires one to filter out the projective transformations before matching. Other examples involve matching outlines or surfaces; see for example Chui & Rangarajan (2000) and an unpublished 2002 Ph.D. thesis from the Technical University of Denmark by L. Pedersen. Here no labelling of points is involved, and we are dealing with a continuous contour or surface rather than a finite number of points. Such problems are not addressed in this paper.

The principal innovations in our approach are the fully model-based approach to alignment, the fact that the model formulation allows us to integrate out the hidden point locations, the prior specification for the rotation matrix, and the Markov chain Monte Carlo algorithm.

一般的摘要只是緒論的子集¹, 這篇文章兩者都要唸, 因為緒論裡說得並不多。它先提出除了 bioinformatic 之外, 還有 image analysis, 但是語焉不詳, 只是要我們去看 Mardia et al. (2003)。

然後直接跳到兩組數據。在有數據的論文裡, 數據最重要, 因此得先介紹。但其後所說的問題 “In this problem, one of the main aims is to take a query active site and find matches to a given database, in some ranking order.” 似乎本文根本不會去碰。因此看起來只是花招。

然後他們提到三篇文章: 1998, 2000 和 2002 的博士論文。但全沒有提這些文章和本

¹但最好要文字有所不同。

文的關係，好像它們只是 image analysis，問題類型雖然類似，所用的技術和本文全然不同的樣子。

最後談到本文的賣點：幾乎全都是技術性的。得到甚麼結論呢？似乎看不出來。緒論裡連 matching matrix 提都沒有提，為甚麼要估計它，估得準是不是等於 matching 就做得好，還是說根本只有一種 matching，好不好我們控制不了，最多只是量測一下 matching 的結果而已？

這是可以去讀的文章，因為它似乎留下很多空間讓我們去問新的問題。此外，相關的文獻似乎不多，這暗示問題也許還新，下功夫搞清楚了就有機會。

關於 Zheng, Salganik and Gelman 的論文

這篇文章刊在 2006 的 *Journal of the American Statistical Association*, **101**, 409-423. 這是美國統計學會旗下的招徠期刊。三位作者都來自 Columbia 大學，其中，Salganik 是社會學的博士生，其他兩位都是統計系的老師。

標題就十分搶眼：“How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks”，光是憑這個標題，很多人就會來讀一下論文摘要²。標題已抓全要點：正面的使用 (using)，而不是將 overdispersion 當作數據中的缺失來小心處理。這樣的態度是值得提的，而標題中將它反應出來。

文章的摘要如下：

Networks — sets of objects connected by relationships — are important in a number of fields. The study of networks has long been central to sociology, where researchers have attempted to understand the causes and consequences of the structure of relationships in large groups of people. Using insight from previous network research, Killworth et al. and McCarty et al. have developed and evaluated a method for estimating the sizes of hard-to-count populations using network data collected from a simple random sample of Americans. In this article we show how, using a multilevel overdispersed Poisson regression model, these data also can be used to estimate aspects of social structure in the population. Our work goes beyond most previous research on networks by using variation, as well as average responses, as a source of information. We apply our method to the data of McCarty et al. and find that Americans vary greatly in their number of acquaintances. Further, Americans show great variation in propensity to form ties to people in some groups (e.g., males in prison, the homeless, and American Indians), but little variation for other groups (e.g., twins, people named Michael or Nicole). We also explore other features of these data and consider ways in which survey data can be used to estimate network structure.

它強調幾點：(1) Social network 的研究，有其歷史；(2) 作者用 “multilevel overdispersed Poisson regression model” 來做分析；(3) 對於估計 hard-to-find 群體大小的方法，是參考了 Killworth et al. and McCarty et al. (注意到，在摘要中不必將論文的年份寫出來)；(4) 利用 variation 是本文技術上的創新，和以前的論文，皆有所不同³；(5) 證明美國人的社交傾向，變異極大；(6) 甚至考慮到如何估計 network structure —— 雖然沒有說

²當然是因為一般的統計論文，不止是內容，連標題都十分乏味。

³其實，正面面對變異，統計學者早已覺醒。如質量問題中的田口方法，還有如 ARCH 之類的時間序列模型都是。

明如何定義一個被估計的 network structure。

這個摘要比較平實。主要是因為 social network 是生活的一部分，即使不知道科學上的定義，也可以有一些感覺。這摘要給我們的印象是，作者也做了不少工作，並且有些工作還是創新的。他們正面地面對 overdispersion，並且有趣地處理了真實數據。

緒論是這樣的 (中文的部分是我加上的):

Recently a survey was taken of Americans asking, among other things, “How many males do you know incarcerated⁴ in state or federal prison?” The mean of the responses to this question was 1.0. To a reader of this journal, that number may seem shockingly high. We would guess that you probably do not know anyone in prison. In fact, we would guess that most of your friends do not know anyone in prison either. This number may seem totally incompatible with your social world.

這一段就開啓得不錯——你的朋友裡有多少在坐牢？平均數是一個人？怎麼這樣高？這樣豈不是一半的人都在牢裡？美國的治安真可怕啊！有了這樣驚人的數據，然後再輕鬆地為你解套：本文的讀者大概都是教育程度好，社經地位較高的一族，多半沒有認識坐過牢的人。於是你心情愉快，好奇心起，這個「平均一人」是從哪裡出來的？

So how was the mean of the responses 1? According to the data, 70% of the respondents reported knowing 0 people in prison. However, the responses show a wide range of variation, with almost 3% reporting that they know at least 10 prisoners. Responses to some other questions of the same format, for example, “How many people do you know named Nicole?” show much less variation.

這一段明白告訴讀者數據中的 variation 極大——有些人認識一大堆坐牢的人⁵。儘管，70% 的人都不認識坐牢的人。

This difference in the variability of responses to these “How many X’s do you know?” questions is the manifestation of fundamental social processes at work. Through careful examination of this pattern, as well as others in the data, we can learn about important characteristics of the social network connecting Americans, as well as the processes that create this network.

前面兩段話都沒有技術用語。但這一段開始使用了 social network 以及它的形成過程——社會科學研究的味道就進來了。

This analysis also furthers our understanding of statistical models of two-way data, by treating overdispersion as a source of information, not just an issue that requires correction. More specifically, we include overdispersion as a parameter that measures the variation in the relative propensities of individuals to form ties to a given social group, and allow it to vary among social groups. Through such modeling of the variation of the relative propensities, we derive a new measure of social structure that uses only survey responses from a sample of individuals, not data on the complete network.

這一段特將 “include overdispersion as a parameter that measures the variation in the

⁴這個字指的是「被監禁」。

⁵當然不是指牢頭。

relative propensities of individuals”加以說明，並申明我們利用 “modeling of the variation of the relative propensities”，才導出 “a new measure of social structure that uses only survey responses”。

Background

Understanding the structure of social networks, and the social processes that form them, is a central concern of sociology for both theoretical and practical reasons (Wasserman and Faust 1994; Freeman 2004). Social networks have been found to have important implications for the social mobility (Lin 1999), getting a job (Granovetter 1995), the dynamics of fads and fashion (Watts 2002), attitude formation (Lee, Farrell, and Link 2004), and the spread of infectious disease (Morris and Kretzschmar 1995).

這一段選用一些關於 social network 的一些文獻，注意到沒有太老的東西。這雖不是一個完整的文獻蒐尋，但至少資料都還新。表示作者還知道最近發生了甚麼⁶。

When talking about social networks, sociologists often use the term “social structure,” which, in practice, has taken on many different meanings, sometimes unclear or contradictory. In this article, as in the article by Heckathorn and Jeffri (2001), we generalize the conception put forth by Blau (1974) that social structure is the difference in affiliation patterns from what would be observed if people formed friendships entirely randomly.

用文字說明本文中所謂的 social structure 是甚麼意思。觀念是相對的：將「隨機交友」當作 social structure 為 0 來看問題。當然，這裡還沒有明確的定義，只有明確的想法⁷。

Sociologists are not the only scientists interested in the structure of networks. Methods presented here can be applied to a more generally defined network, as any set of objects (nodes) connected to each other by a set of links (edges). In addition to social networks (friendship network, collaboration networks of scientists, sexual networks), examples include technological networks (e.g., the internet backbone, the world-wide web, the power grid) and biological networks (e.g., metabolic networks, protein interaction networks, neural networks, food webs); reviews have been provided by Strogatz (2001), Newman (2003b), and Watts (2004).

這一段是可有可無的。只是說「可以有更廣的應用」，一般人都不會去查的，但作者應該真的讀過所列的幾篇文章⁸，否則就是不負責任。

Overview of the Article

In this article we show how to use “How many X’s do you know?” count data to learn about the social structure of the acquaintanceship network in the United States. More specifically, we can learn to what extent people vary in their number of acquaintances, to what extent people vary in their propensity to form ties to people in specific groups, and also to what extent specific subpopulations (including those that are otherwise hard to count) vary in their popularities.

⁶引用十年以前的文章時，最好只引用那些經典的，才顯得你有學問有品味。

⁷如果你有相關的數據，可考慮做「論網路交友的社會結構」。

⁸至少他也得用細讀過它們的緒論。

先提出數據的基本形態：由某一特殊問題 “How many X’s do you know?” 而得，所以是 count data。

The data used in this article were collected by McCarty, Killworth, Bernard, Johnsen, and Shelley (2001) and consist of survey responses from 1,370 individuals on their acquaintances with groups defined by name (e.g., Michael, Christina, Nicole), occupation (e.g., postal worker, pilot, gun dealer), ethnicity (e.g., Native American), or experience (e.g., prisoner, auto accident victim); for a complete list of the groups, see Figure 4 in Section 4.2. Our estimates come from fitting a multilevel Poisson regression with variance components corresponding to survey respondents and subpopulations and an overdispersion factor that varies by group. We fit the model using Bayesian inference and the Gibbs-Metropolis algorithm, and identify some areas in which the model fit could be improved using predictive checks. Fitting the data with a multilevel model allows separation of individual and subpopulation effects. Our analysis of the McCarty et al. data gives reasonable results and provides a useful external check on our methods. Potential areas of further work include more sophisticated interaction models, application to data collected by network sampling (Heckathorn 1997, 2002; Salganik and Heckathorn 2004), and application to count data in other fields.

這一段說明數據的來源 (McCarty, Killworth, Bernard, Johnsen, and Shelley (2001)) 和一般說明。所用到的模型 (a multilevel Poisson regression with variance components), 所用的工具 (Bayesian inference and the Gibbs-Metropolis algorithm) 及甚麼時候可以做得更好 (identify some areas in which the model fit could be improved using predictive checks)。至於這對於 social network 研究的影響，倒是看不出來。

主要的結果似乎仍然是導出了 “a new measure of social structure that uses only survey responses”。這對於社會學的意義是甚麼？似乎沒有著墨，甚至不知道這個所謂的 new measure，是用來量測甚麼性質的。

Buhlmann 的文章

這文章發表在 2006 的 Annals of Statistics, 34, 559-581. Annals 是國際數理統計學會的旗艦期刊。要求證明嚴謹，是技術性要求很高的。當然，這篇文章是蠻理論的 (否則不會投進來)。讀它的摘要和緒論，除非你是相當夠格的專家，根本就搞不清來他做了甚麼工作。我是一直讀到 p.565 才把此文的「成就」搞明白⁹。但還不算數學證明。

這是為一小撮人寫的論文。標題是 “Boosting for high-dimensional linear models”。這裡只用了一個較生疏的名詞：boosting。這個名詞來自所謂的機械學習 (machine learning)，原意是指將一個較弱的 (預測) 方法，經過某類處理之後 (處理的方式叫 boosting)，變得更好¹⁰。

統計學家對 boosting 有興趣，因此也有不少學者加入研究。但此文光是從標題是看不出甚麼的，所以先讀一下它的摘要：

We prove that boosting with the squared error loss, L_2 Boosting, is consistent for very

⁹因為好幾個名詞到後面才明確定義。這對於摘要和緒論是可以的，對正文就不行。對這類文章，作者並沒有希望你第一次就讀懂。這種摘要和緒論，是為讓專家留下深刻印象用的。

¹⁰我曾以為如果我有 51% 的機率猜到明天某股票的漲或跌，則我應可用 boosting 的辦法，提高我的猜測率到 90%，boosting 當然沒有這樣利害。

high-dimensional linear models, where the number of predictor variables is allowed to grow essentially as fast as $O(\exp(\text{sample size}))$, assuming that the true underlying regression function is sparse in terms of the ℓ_1 -norm of the regression coefficients.

這個摘要太專業了，我要從論文的 p.559 讀到 p.565 才知道它第一句在說甚麼。基本上是這樣的。在

$$y = X\beta + \epsilon, \quad (y = n \times 1, X = n \times p, \beta = p \times 1)$$

這樣的線性模型中，一般都要要求 $n > p$ ，因為否則「自由度不夠」。但如果 $p > n$ ，請問如何估計 $E(y|X)$ 這個迴歸函數？一般當然是不行，否則我們讀到的初等統計都錯了。但是，如果大多數的 β 的值都很小時是不是有可能？

這裡，「大多數的 β 的值都很小」的條件，叫做 sparse。這篇文章所證明的是，在 sparse 的條件下，是可以得到 $E(y|X)$ 的「一致估計量 (consistent estimate)」的。

In the language of signal processing, this means consistency for de-noising using a strongly overcomplete dictionary if the underlying signal is sparse in terms of the ℓ_1 -norm. We also propose here an AIC-based method for tuning, namely for choosing the number of boosting iterations. This makes L_2 Boosting computationally attractive since it is not required to run the algorithm multiple times for cross-validation as commonly used so far.

這一段說，在做 boosting 時，有一個「每次疊算的次數」要交待明白。作者說他有辦法。

We demonstrate L_2 Boosting for simulated data, in particular where the predictor dimension is large in comparison to sample size, and for a difficult tumor-classification problem with gene expression microarray data.

這一段對數理統計的論文是蠻標準的，先講方法，再做模擬，最後再套入真實數據。這裡，真數據來自生物微晶片，雖然不是作者自己測出的，但這個問題仍然紅火（事實上，我們會去對 $n < p$ 的情形如此關注，這類數據的大量存在是主要原因）。

這個摘要寫得算中肯——反正只是寫給內行人看的——它只是明白地講出所做出的主要結果。

它的緒論則較長。其實只是摘要的加強版。內容如下：

Freund and Schapire's [11] AdaBoost algorithm for classification has attracted much attention in the machine learning community (cf. [20] and the references therein) as well as in related areas in statistics [1,13], mainly because of its good empirical performance with a variety of datasets. Boosting methods were originally introduced as multiple prediction schemes, averaging estimated predictions from reweighted data. Later, Breiman [1, 2] noted that the AdaBoost algorithm can be viewed as a gradient descent optimization technique in function space. This important insight opened a new perspective, namely to use boosting methods in contexts other than classification. For example, Friedman [12] developed boosting methods for regression which are implemented as an optimization using the squared error loss function: this is what we call L_2 Boosting. It is essentially the same as Mallat and Zhang's [19] matching pursuit algorithm in signal processing.

這一段是給關於 boosting 的背景資料。主要是說誰做了甚麼工作。因為指明了 L_2 Boosting 是本文的手段 (先別管 L_2 Boosting 的定義, 對這類數理統計的文章, 早晚會給出明確定義的), 此段只誇獎了 [1, 2] 兩篇文章, 說 “This important insight opened a new perspective”。至於其它的論文, 都不帶褒貶。他並提出 L_2 Boosting 和 match pursuit 是同一回事¹¹。注意到, 他這一段只提了七篇論文, 這都是有選擇的, 這些都是關於 L_2 Boosting 的重要文獻, 充分表示作者是否內行。

Recently, Efron, Hastie, Johnstone and Tibshirani [10] made a connection for linear models between forward stagewise linear regression (FSLR), which seems closely related to L_2 Boosting, and the ℓ_1 -penalized Lasso [22] or basis pursuit [5]. Roughly speaking: under some restrictive assumptions on the design matrix of a linear model, FSLR approximately yields the set of all Lasso solutions (when varying over the penalty parameter). This intriguing insight may be useful to get a rough picture about L_2 Boosting via its relatedness to FSLR: it does variable selection and, shrinkage, similar to the Lasso. However, it should be stated clearly that the methods are not the same; an example showing a distinct difference between L_2 Boosting and the Lasso is presented in Section 4.3. Moreover, we point out in Section 2.1 that FSLR and L_2 Boosting are different algorithms as well.

這一段又提了一些相關的論文, 說它們似乎和 L_2 Boosting 做同樣的事。作者的結論反面的: 這些都不等於 L_2 Boosting。他明白地說: “it should be stated clearly that the methods are not the same”。這一段所給人的印象是: 此人將 L_2 Boosting 的事情已弄得十分清楚了。也澄清了一些大家或以為對的事情。而這些都是學問, 新的知識。

As the main result, we prove here that L_2 Boosting for linear models yields consistent estimates in the very high-dimensional context, where the number of predictor variables is allowed to grow essentially as fast as $O(\exp(\text{sample size}))$, assuming that the true underlying regression function is sparse in terms of the ℓ_1 -norm of the regression coefficients.

這裡強調本文的主要結果。

This result is, to our knowledge, the first about boosting in the presence of (fast) growing dimension of the predictor.

並且說明別人沒有做出這類型態的結果 —— 該說自己好的時候就大方地說。

Some consistency results for boosting with fixed predictor dimension include [17, 18] as well as [25]. Except for Jiang's [17] result, these authors consider versions of boosting either with ℓ_1 -constraints for the boosting aggregation coefficients or, as in [25], with a relaxed version of boosting which we found very difficult to use in practice due to the nonobvious tuning of the relaxation, that is, how fast the boosting aggregation coefficients should decay. The result by Zhang and Yu [25] may be generalized without too much effort to a setting with increasing dimension of the predictor variable, but their theoretical work includes only a rigorous treatment of the classification problem (besides the above mentioned disadvantage of their relaxed boosting algorithm).

這裡說其它似乎相近的結果都不夠好。注意到作者的理由都很明確, 不是空口說白話的。需知這樣的內容最易讓同儕生氣。文字不要帶火氣。

¹¹這是暗示「本人學問不錯」的手法, 要學 —— 尤其是他不僅提 [19], 還要提 “signal processing”, 這秀出了「格局不止在 machine learning」。

We believe that it is mainly for the case of high-dimensional predictors where boosting, among other methods, has a substantial advantage over more classical approaches. Some evidence for this will be given in Section 4.1, and other supporting empirical results have been reported in [3] in the different context of low- or high-dimensional additive models for comparing L_2 Boosting with more traditional methods such as backfitting or MARS (restricted to additive function estimates). Notably, many real datasets nowadays are of high-dimensional nature. Besides the well-documented good empirical performance of boosting, we identify it here as a method which can consistently recover very high-dimensional, sparse functions.

此處說明高維的問題才是主要的應用，並且因為有 “a substantial advantage over more classical approaches”，世上有太多這類的數據（這一點其實大家都知道）。並且，最主要的，作者又將主結果再提一次。

We may also view our result as a consistency property for de-noising using L_2 Boosting with a strongly overcomplete dictionary. In contrast to a complete dictionary, for example, Fourier- or wavelet-basis, the strongly overcomplete noisy case is not well understood. Our result yields at least the basic property of consistency.

這一段，作者又將本文與其它事物拉關係。這類問題的結構大同小異，多半在不同的架構下運作。作者說 “Our result yields at least the basic property of consistency”，但這一段話有「想當然耳」的味道——這些東西是可以給學生做的題目，有作者這樣的專家猜應該會成立，但真正寫出來卻是另一件事。

Besides the theoretical consistency result, we propose here a computationally efficient approach for the tuning parameter in boosting, that is, the number of boosting iterations. We give an easily computable definition of degrees of freedom for L_2 Boosting, and we then propose its use in the corrected AIC criterion. Unlike cross-validation, our AIC-tuning does not require boosting to be run multiple times. This makes the AIC-type data-driven boosting computationally attractive: depending on the data, it is sometimes as fast as the very efficient LARS algorithm for the Lasso with tuning by its default ten-fold cross-validation [6, 10].

這一段說明計算的效率。作者也有做法：用 AIC 來做。到此為止，論文的理論部分都已介紹完畢。此文其實只有一個主定理，但作者洋洋灑灑地用了 25 頁的 Annals pages，主要是靠此人確實對相關的問題知道得夠多夠深入，所以才寫得出這些旁徵博引的話。需知說這類話最易出問題，因為你的看法不見得就是評審人的看法。這時深度就變得重要，如果你真的言之有物，那些專家也會懂的。

We demonstrate on some simulated examples how our L_2 Boosting performs for (low-and) mainly high-dimensional linear models, in comparison to the Lasso forward variable selection, ridge regression, ordinary least squares and a method which has been designed for high-dimensional regression [14]. We also consider a difficult tumor-classification problem with gene expression microarray data: the predictive accuracy of L_2 Boosting is compared with four other, commonly used classifiers for microarray data, and we briefly indicate the interpretation of the L_2 Boosting fit along the lines of a linear model fit.

最後說明模擬和真實數據的套用，以及與其它類似方法的比較。

這篇論文的摘要和緒論都算是標準的寫法，它沒有奇兵突出（像 Zheng et al. 一樣），但平鋪直敘，立刻進入主結果，並大力地和現有的文獻比較，以宣揚自己的優點，坐實別人的缺點。這是數理統計論文的通常寫法。

Brien 和 Bailey 的論文

這篇文章登在 *Journal of the Royal Statistical Society, Series B* (2006) **68**, 571-609. 這是英國皇家統計學會的招牌刊物。

標題極短：“Multiple randomization”，只有兩個字。一般只有自認是高手者才會這樣取論文標題。這暗示本文並不針對某一特定的小範圍而討論。它的摘要是要這樣的：

Multitiered experiments are characterized by involving multiple randomizations, in a sense that we make explicit. We compare and contrast six types of multiple randomizations, using a wide range of examples, and discuss their use in designing experiments. We outline a system of describing the randomizations in terms of sets of objects, their associated tiers and the factor nesting, using randomization diagrams, which gave a convenient and readily assimilated summary of an experiment's randomization. We also indicate how to formulate a randomization-based mixed model for the analysis of data from such experiments.

這裡用了 “tier” 一詞，這並不是一般的實驗設計 (DOE) 教本上常見的。此處就特用來和 randomization 一起配合使用（所以後文將再說明）。摘要裡說明本文是要用很多例子來討論多重隨機化的道理的，提到了 “randomization diagrams”，和如何處理 “randomization-based mixed model”。這個摘要沒有甚麼不對，只是看不太懂而已。DOE 已是非常精微細緻的學問，細緻到非專家就不想去碰的程度。作者之一的 Bailey 是此中的有名的專家（如果不是權威），但這樣的摘要，我讀起來還是有「閑人免進」的感覺。

它的緒論如下：

Experiments are distinguished from observational studies and happenstance data by the purposive application of treatments to observational units. Nelder (1965a, b), White (1975), Bailey (1981, 1991) and Heiberger (1989) formulated methods for the analysis of experiments that take this distinction into account by classifying factors in the experiment as either ‘block’ or ‘treatment’ factors. Here we interpret block and treatment factors to mean the sets of unrandomized and randomized factors respectively. Many researchers have advocated the use of such a distinction (Fisher (1935), Wilk and Kempthorne (1957), Cox (1958), section 6.3, Yates (1975), Mead and Curnow (1983), section 14.1, Brien (1983) and Piepho et al. (2003)). The distinction enables the direct construction of analysis-of-variance tables exhibiting all the confounding, as well as the inclusion, in a mixed model, of all terms that are warranted by the randomization.

這一段是名家手筆。把區集 (block) 和 (treatment) 用「不隨機化」和「隨機化」來分辨。等閑人寫不出來。

Of course, it is possible to formulate an analysis without making this distinction. For example, in the analysis of a randomized complete-block design as a ‘two-way analysis of variance’ based on a two-factor, no-interaction model, no distinction is made between the treatments that are randomly assigned and the blocks that result from inherent features of the observational units. As Kempthorne (1955) noted, this results in the same analy-

sis for the randomized complete-block design and the two-way factorial experiment, even though they differ markedly in the randomization that is employed. This is because the confounding that is inherent in each experimental situation has not been recognized (Brien and Payne, 1999). Similarly, the common description of the classic split-plot experiment in terms of only the three factors blocks, mainplot treatments and subplot treatments also fails to distinguish between randomized treatments and inherent features of the observational units.

讀過一點 DOE 的人都知道 “this results in the same analysis for the randomized complete-block design and the two-way factorial experiment, even though they differ markedly in the randomization that is employed”, 但讀過 Brien and Payne (1999) 並肯去深入了解的人就少了。若是能夠了解到 “the confounding that is inherent in each experimental situation has not been recognized”, 是否實驗數據的分析方法就會不同?

Brien (1983) identified experiments that involve more than the single randomization of treatments to plots and concluded that two sets of factors are not enough to describe the randomization in such experiments. The general term ‘tiers’ was introduced for the sets of factors in an experiment which result from the classification of the factors according to their status in the randomization: see also Cullis et al. (2003). Experiments that involve multiple randomizations, and hence more than two tiers, were labeled ‘multitiered’. Multitiered experiments include two-phase, some superimposed and some single-stage experiments, and some multistage experiments using the same units at each stage; they do not represent a collection of new designs, but are a class of designs made up of several existing design types.

這一段說明, 在 DOE 中, “experiments that involve more than the single randomization of treatments” 的情形, 20 年前就已發現。並由此引入 “tier” 一詞, 並讓它和「隨機化」相連。照顧到摘要開始時的句子。但像 “Multitiered experiments include two-phase, some superimposed and some single-stage experiments, and some multistage experiments using the same units at each stage” 的寫法, 就太專業了, 我讀了一點感覺都沒有。也許可以 impress 專家罷。

Two-phase experiments were introduced by McIntyre (1955) and were discussed by Cox (1958), although Cox used the term ‘stage’ rather than ‘phase’. They are characterized by the following:

- (a) a complete experiment in the first phase, although not necessarily with the measurement of response variables;
- (b) the randomization of the units from the first phase to the units in the second phase.

到此處, 過去的和本文有關的文獻回顧就差不多結束了。這幾段算是較長的, 但是, 懂就懂; 不懂還是不懂。因為搞懂也需要功力。至於文字, 是寫得有專家的味道。

A common situation in which they occur is where the produce from a field trial has to be processed in a laboratory or an evaluation phase (Brien, 1983; Brien et al., 1987; Wood et al., 1988; Brien and Payne, 1999; Cullis et al., 2003; Kerr, 2003). All of the examples of two-phase experiments that have been published in the literature employ only one of the simplest type of multiple randomization.

In this paper we distinguish between stages and phases in experiments, the former including the latter as a special type. We note that usage of stages and phases in experiments is unrelated to that in sampling. Multistage experiments are conducted in several distinct time intervals with randomization of factors for each interval. They include two-phase, superimposed and change-over experiments (but not longitudinal studies), as well as multistage reprocessing experiments, in which the response is measured only after several stages of processing the same units (Miller (1997), Mee and Bates (1998) and Box and Jones (1992), section 5). Not all multistage experiments are multitiered as they do not involve multiple randomizations — the usual repeated measurements experiments that are discussed in text-books are not. Single-stage experiments that involve multiple randomizations include some grazing and some plant experiments. For example, Brien and Demetrio (1998) discussed multiple randomizations in the context of continuous grazing trials and Preece (1991) described a multitiered horticultural experiment.

由此段開始介紹本文所要做的工作。先說明本文中所用到 “stage” 和 “phase” 意味上的不同。

Although the experiments do not involve new designs, there has been uncertainty and difficulty in their randomization and analysis. The purpose of this paper is to describe and compare six different types of multiple randomization and how they are employed in a range of multitiered experiments. We concentrate on equireplicate designs, because they demonstrate all the relevant issues but without extra complications. We hope to increase understanding of the way in which multiple randomizations can be employed in designing experiments. We also develop randomization diagrams for depicting the randomization in an easily absorbed form. In Section 2, we introduce and define the terms that are required and illustrate their use in the context of two-tiered experiments. Section 3 describes a simple three-tiered experiment. The six types of multiple randomizations are then described: Section 4 discusses and gives examples of multiple randomizations in which the two constituent randomizations can be performed in either order, whereas Section 5 does the same for the inclusive multiple randomizations. Section 6 gives examples involving more than one type of multiple randomizations, and hence three or more randomizations. A strategy for formulating a mixed model for analyzing multitiered experiments is outlined in Section 7. Section 8 addresses some general issues concerning terminology and multiple randomizations. Further examples involving multiple randomizations are available in Brien (1992) and from the multitiered experiment Web site at <http://chris.brien.name/multitier/>.

這一段說本文所做的事情，是比摘要要說得多的完整版。所做的事有二：“to describe and compare six different types of multiple randomization and how they are employed in a range of multitiered experiments” 和 “develop randomization diagrams for depicting the randomization in an easily absorbed form”。最後，再用「目錄式」的列舉，告訴讀者在哪一節做哪一件事。這是緒論的習慣，雖然，我每次唸到此就，就必然會跳過。

大體說來，這仍然是一篇中規中矩的論文緒論——不能因為我對 DOE 知識淺薄而小看它。因為它的目的，不是讓我覺得這文章好，而是讓 DOE 的專家覺得這文章內容。

這是一篇所謂的 discussion paper——原則上應該是較有份量的文章。統計學者常以「被邀往 RSS 宣讀（並討論）論文」為榮。

題外的感覺

The relative performance of the tests are assessed by applying them to simulated data and comparing their size and power; one of the modified tests (the CW test) is shown to be superior.

Introduction

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive method that produces a time sequence of images of a subject's brain that are sensitive to changes in blood oxygenation caused by neural activation. The vast majority of analytical techniques that are applied to fMRI data assume the transfer function between neural activation and subsequent changes in blood oxygenation, the haemodynamic response function (HRF), is known fully and the data follow the Gaussian distribution. In this article, we consider the analysis of fMRI data collected in one of two states, called "activation" and "rest", based on two-sample tests. From knowledge of the length of the haemodynamic delay, measurements during the transition period between activation and rest can be omitted. The validity of the classical two-sample t-test is based on the assumption that the activation data and the rest data are Gaussian with equal variances. In this article, we propose use of a modified two-sample test for fMRI data that allows for departures from this assumption. We study three competing tests. One is the Welch test (Welch, 1937) which is a modification of two-sample t-test that allows unequal covariances. A second competitor is the Cressie-Whitford (CW) test (Cressie and Whitford, 1986) that can be used with non-Gaussian data. The third competitor is the Wilcoxon rank (WR) test (Wilcoxon, 1945). In what follows, we compare the classical t-test with the Welch, CW, and WR tests for fMRI data based on a block design, where the blocks alternate between periods of activation and rest.

The next section describes the physiological background and physical processes used in fMRI and the most common methods used to process fMRI data; it also defines the four two-sample tests (including the classical two-sample t-test) that are compared in Section 4. Section 3 discusses the application of the two-sample tests for fMRI data and describes the methods used to identify and quantify departures from Gaussianity for each voxel. The size and power of the four tests are compared in Section 4 using a simulation study of fMRI data, from which recommendations are given. Section 5 contains discussion and conclusions.

第二篇

An Evaluation of Multiple Behavioral Risk Factors for Cancer in a Working Class, Multi-Ethnic Population

Abstract

Behavioral risk factors for cancer tend to cluster within individuals, which can compound risk beyond that associated with the individual risk factors alone. There has been increasing attention paid to the prevalence of multiple risk factors (MRF) for cancer, and to the importance of designing interventions that help individuals reduce their risks across multiple behaviors simultaneously. The purpose of this paper is to develop methodology to identify an optimal linear combination of multiple risk factors (score function) which would facilitate evaluation of cancer interventions.

Introduction

Despite the considerable biomedical advances of the last half-century, facilitating improvement in lifestyle behaviors remains the most efficacious population-level strategy for reducing cancer risk. Estimates vary, but suggest that over fifty percent of new cancer cases and up to one-third of cancer mortality could be prevented through improvements in health behavior practices (American Cancer Society, 2004; Doll and Peto, 1981). A 19 percent decline in the rate at which new cancer cases occur, and a 29 percent decline in the rate of cancer deaths, could potentially be achieved by 2015, if prevention efforts were heightened and behavior change sustained. This would translate to the prevention of approximately 100,000 cancer cases and 60,000 cancer deaths each year, by the year 2015 (National Cancer Policy Board and Institute of Medicine, 2003). There is ample epidemiological evidence for the consideration of red meat consumption, physical activity, and folic acid intake in cancer prevention efforts. Regular physical activity lowers the risk of cancers of the colon, breast, and possibly prostate (Colditz, Cannuscio, and Frazier, 1997; Friedenreich and Rohan, (1995)). An additional 30 percent of cancer deaths can be attributed to adult diet (Anonymous, 1996); higher intake of red meat has been associated with increased risk of colon (Sandhu, White and McPherson, 2001) and prostate cancers (Michaud, Augustsson, Rimm, Stampfer, Willett, and Giovannucci 2001). Associated with both physical inactivity and diet is obesity, which may account for between 25-30 percent of cancers of the colon, breast (postmenopausal), endometrium, kidney, and esophagus (Vainio and Bianchini, 2002). Folic acid is protective against colon cancer (Giovannucci *et al.* 1998); long-term multi-vitamin use, in particular has been found to reduce risk for colon cancer, likely because of its folic acid content (Giovannucci *et al.* 1998). The risk for many diseases, including colon cancer, is associated with multiple behavioral risk factors (MRF); these behaviors are highly interrelated and tend to cluster within individuals. For example, those who eat high-fat diets are also more likely to be sedentary, suggesting that the behaviors may be mutually reinforcing (see e.g., Emmons, Marcus, Linnan, Rossi, and Abrams, 1999). Change in one behavioral risk factor thus may serve as a stimulus or gateway for change in the other health behaviors (see e.g., Emmons *et al.* 1999), and there are overarching behavioral principles and intervention frameworks that guide behavior change efforts across risk factors. Consequently, to facilitate population-level reductions in cancer risk, it may be inefficient to target discrete behavioral risk factors, when similar principles might be applied simultaneously to multiple behaviors (Institute of Medicine, 2000). The literature provides little consensus as to the most appropriate analytic strategy for evaluating the efficacy of MRF interventions; most studies have analyzed the various outcomes independently or by creating a simplistic sum (e.g., 1 RF + 1 RF = 2RFs) (see e.g., Prochaska and Sallis 2004; Campbell, James, Hudson, Carr, Jackson, Oakes, Demissie, Farrell, and Tessaro, 2004). This could be problematic, because the use of separate analytic strategies may result in improper inferences regarding the effect of an MRF intervention because of correlation among the factors. Such strategies may overlook the clustering effect brought about by the agglomeration of multiple behavioral risk factors and have been criticized as being too simplistic. The purpose of this paper is to develop a methodology to identify an optimal linear combination of multiple behavioral risk factors (MRF score function) for cancer that would best facilitate evaluation of an MRF cancer intervention.

第三篇

Reducing Subjectivity in the Likelihood

Abstract

Some scientists prefer to exercise substantial judgment in formulating a likelihood function for their data. Others prefer to try to get the data to tell them which likelihood is most appropriate. We suggest here that one way to reduce the judgment component of the likelihood function is to adopt a mixture of potential likelihoods and let the data determine the weights on each likelihood. We distinguish several different types of subjectivity in the likelihood function and show with examples how these subjective elements may be given more equitable treatment.

Introduction

We propose methods for modeling the likelihood function that will require fewer subjective judgments. We first discuss the nature of the problem of subjectivity in the likelihood function; then we review some related research; and finally, we define a mixture likelihood function and suggest estimation procedures that reduce the effects of subjective views imposed on the observed data.

1.1 Statement of the Problem

It is sometimes desirable that beliefs of experimenters should be brought into a scientific analysis in ways that minimally distort the measured data (see, for example, Hogarth, 1980; Kyberg and Smokler, 1980; Lad, 1996). But that having been said, scientists observing data sometimes interpret the data points subjectively, according to what they want the data to show, and according to how precisely they believe the data points were measured. The latter procedure is of course quite common. This subjective interpretation of observed data may be totally at the unconscious level, or it may be purposeful (with the purposeful interpretation, the analysis may become fraudulent; see for example, Grayson, 1995, 1997; Howson and Urbach, 1990; and Press and Tanur, 2001).

The subjective interpretation of empirical data in medicine was discussed by Kaptchuk (2003). He stated (page 1, *op. cit.*):

“Doctors are being encouraged to improve their critical appraisal skills to make better use of medical research. But when using these skills, it is important to remember that interpretation of data is inevitably subjective and can itself result in bias. Facts do not accumulate on the blank slates of researchers’ minds, and data simply do not speak for themselves. Good science inevitably embodies a tension between the empiricism of concrete data and the rationalism of deeply held convictions. Unbiased interpretation of data is as important as performing rigorous experiments. This evaluative process is never totally objective or completely independent of scientists’ convictions or theoretical apparatus”.

Statistical analysis of a data set most often proceeds by summarizing the distribution of the data in terms of its likelihood function. In order to specify the form of the likelihood function, various assumptions are made about the data, such as mutual independence, identical distributions, unimodality, etc. After the likelihood function has been specified, additional assumptions are sometimes made (significance levels thought to be appropriate are specified, a prior distribution about the underlying unobservable quantities may be

brought in, etc.). Analysis of the data generally proceeds by trying to keep the likelihood function treatment of the data as simple as possible, so that the scientist or analyst will introduce minimal distortion of the data. The analyst tries not to discard data, and tries to maximize the chance of understanding what nature is trying to tell us through the revealed data about the underlying phenomenon. In this way, when the analysis of the data has been completed, the claim can reasonably be made that the conclusions drawn from the analysis approximate, if not precisely reflect, the laws of nature, rather than the possible misinterpretations and misunderstandings of the laws of nature by human beings. It will be useful to first briefly define what we mean by objectivity and subjectivity, in this context.

According to Mandik, 2001,

“The word objectivity refers to the view that the truth of a thing is independent from the observing subject. The notion of objectivity entails that certain things exist independently from the mind, or that they are at least in an external sphere. Objective truths are independent of human wishes and beliefs. The notion of objectivity is especially relevant to the status of our various ideas, and the question is to what extent objectivity is possible for thought, and to what extent it is necessary”.

This is but one of many definitions that have been suggested. The elusive quest for objectivity in science has been, and remains, an important topic of discussion among historians and philosophers of science (for extensive additional discussions of the meaning of “objectivity”, see for example, Bower, 1998; Porter, 1995, 1996; and Daston and Galison 1992). For some, scientific objectivity involves the search for certainty in knowledge about one of nature’s well-kept secrets, independent of what human beings believe; but in many cases, we find that what we earlier thought to be true about nature, turns out later to be questionable.

In an interesting example from physics, Folger, 2003, pointed out that:

“Pioneer 10, launched in 1972, is now some 8 billion miles from home. But is has been slowing down, as if the gravitational pull on it from the sun is growing progressively stronger the farther away it gets. Milgrom proposed (see the MOND pages-MODified Newtonian Dynamics) that Newton’s laws might change at these accelerations. If Milgrom is right, Newton’s and Einstein’s laws will be in for some major tweaking.”

Sometimes the scientist has such deep understanding and insight into the phenomenon he/she is studying that the scientist’s own predictions of what should be found from the analysis are far superior to what the data analysis seems to indicate. In some cases the beliefs of the scientist or analyst are so strong, even before actually taking any data that bear on the phenomenon, that the data are interpreted or manipulated so that they will reflect these preconceived views of the scientist. Any preconceived personal views (views held before taking any data), weak or strong, are what we refer to in this context as subjectivity.

1.2 Related Research

One approach to reducing the effects of differing assumptions about likelihoods may

be found in a line of research that involves use of the empirical likelihood function. In this approach, most useful in large samples, a discretized, binned, version of the empirical cdf, instead of a specific likelihood function, is used. Inference is then made from a multinomial distribution. An unfortunate feature of this approach is the additional unknown parameters that are concomitantly introduced into the model. See: Owen, 1988, 2001. For typically small and moderate size samples this could be a problem, but for the massive data sets typical of data mining applications (see, for example: Berry & Linoff (1997); and Hastie, Tibshirani, and Friedman, 2001) such an approach could be a helpful alternative.

We show in the next section how we might understand and account for some types of subjectivity that sometimes enters the likelihood function, and might not be desired. We will use the definition and form of the likelihood function in which for absolutely continuous random variables, up to a proportionality constant, it is the joint probability density function of the observables given the unobservables.

第四篇

Application of One Sided t-tests and a Generalized Experiment Wise Error Rate, GFWER(k), to High-Density Oligonucleotide Microarray Experiments: An Example Using Arabidopsis

Abstract

Motivation: A formidable challenge in the analysis of microarray data is the identification of those genes that exhibit differential expression. The objectives of this research were to examine the utility of simple ANOVA, one sided t tests, natural log transformation, and a generalized experiment wise error rate methodology for analysis of such experiments. As a test case, we analyzed a Affymetrix GeneChip microarray experiment designed to test for the effect of a CHD3 chromatin remodeling factor, PICKLE (PKL), and an inhibitor of the plant hormone gibberellin (GA), on the expression of 8256 Arabidopsis thaliana genes.

Results: The GFWER(k) is defined as the probability of rejecting k or more true null hypothesis at a given p level, and ignores the probability of making less than k Type I errors. The method was shown to be simple to apply and greatly increases power. A k value as small as 2 or 3 was concluded to be adequate for large or small experiments respectively. A one sided t-test along with GFWER(2)=.05 identified 43 genes as exhibiting PKL-dependent expression. Expression of all 43 genes was re-examined by qRT-PCR, of which 36 (83.7%) were confirmed to exhibit PKL-dependent expression.

Introduction

The advent of inexpensive microarray technology has enabled individual laboratories to easily obtain a global perspective on the expression pattern of thousands of genes. This powerful technology has allowed investigators to diagnose early cancers (Kim, J.W. and Wang, X.W., 2003; Zhang, H. et al., 2003), discover genes that contribute to quantitative traits (Gu, C.C. et al., 2002), and detect coordinated gene regulation during pivotal developmental events such as embryogenesis and sexual maturation (Girke, T. et al., 2000; Lo, J. et al., 2003; Ruuska, S. A. et al., 2002).

The first generation microarrays were generally based on two dye methodologies. These cDNA microarray experiments involve hybridizing two mRNA samples, each of which has been converted into cDNA and labelled with its own fluorophore, on a single glass slide that has been spotted with 10,000-20,000 cDNA probes (Yang and Speed, 2002). In contrast, more recent high-density oligonucleotide microarrays, such as those offered by Affymetrix, provide direct information about the expression levels in an mRNA sample and can have a much higher density (Yang and Speed, 2002).

The majority of methodologies for microarray analysis have been developed for two dye spotted arrays (Kerr et al. 2000; Kerr and Churchill, 2001; for review see Quackenbush, 2001 and Yang and Speed, 2002). Unfortunately these two-dye spotted arrays also pose other statistical issues, such as normalization to correct for dye bias. Furthermore if more than 2 treatments are used, it is not possible to compare all treatments on the same chip thus necessitating an Incomplete Block Design (IBD) type design (Kerr and Churchill, 2001). As such, special experimental designs, such as the reference and rotational design are needed for correct analysis (Kerr and Churchill, 2001; Quackenbush, 2001 and Yang and Speed, 2002).

In contrast, oligonucleotide microarrays use a single dye technology and pose some advantages, including a greatly increased density of genes and simplified experimental design because treatment effects are tested independently on each chip, eliminating the need for IBD designs. Nevertheless, statistical issues remain, such as normality of residuals, homogeneity of residual variance, correlation of errors within an array, and correlation of biological samples across arrays.

Mixed model methods for analysis of microarray experiment, proposed by Wolfinger et al. (2001), solves most of these issues (see Craig et al., 2003 for review). However, the complexity of analysis dramatically increases with these advanced methods. Unfortunately, many of the current practitioners of microarray technology do not possess the mathematical expertise necessary to meaningfully employ these methods. On the other hand ANOVA is a tool that is easy to implement with methods common to most researchers. Kerr and Churchill (2001) conclude that "The analysis of variance (ANOVA) is a natural tool for studying data from experiments with multiple categorical factors".

The first objective of this research was to examine the utility of simple ANOVA for analysis of replicated oligonucleotide microarrays experiments. The motivation was given eloquently by Kerr and Churchill (2001) who stated "An advantage of model based data analysis such as ANOVA is that a model helps the analyst explore the data. If one finds a model inadequate, discovering why it is inadequate can help the analyst identify sources of variation and bias." A secondary objective of this study was to show how using a one sided t-test can be used to increase power. The final objective was to introduce an alternative method to increase power by accepting a base number of false positive with high probability.

The ANOVA is particularly suited to analyzing data from microarray experiments that employ a replicated factorial arrangement of treatments. An example of such an experimental design is one in which the investigator looks at gene expression in wild-type and mutant plants in the presence or absence of an added chemical. Many microarray studies incorporate this type of experimental design, e.g. the response of genes in non-

tumorigenic and tumorigenic tissues to different concentrations of toxic or therapeutic drugs (Lundquist, H. et al., 2002; Martinez, J.M. et al., 2002) or the response of genes from different tissues to estrogen or other hormones (Abe, H. et al., 2003; Faccioli, P. et al., 2002; Fujita, N. et al., 2003; Goda, H. et al., 2002). This design easily extends into any number of genotypes (or tissues) by any number of developmental time points (or biochemical exposures).

The primary biological objective of this research was to understand how a CHD3-chromatin remodeling factor, PICKLE (PKL), and a plant growth regulator, gibberellin (GA), regulate gene expression during germination of Arabidopsis seeds (Rider, S.D. et al., 2003). PKL is necessary for repression of embryonic traits in Arabidopsis (Ogas, J. et al., 1997). Expression of the embryonic state in *pkl* seedlings is inhibited by the plant growth regulator gibberellin (GA) and is enhanced by application of uniconazole-P, an inhibitor of GA biosynthesis (Izumi, K. et al., 1985; Ogas, J. et al., 1997). Specifically, gene expression was examined in wild-type and *pkl* seeds grown in the absence and presence of 10-8 M uniconazole-P. Thus the genotypes were 'wild type' vs. the *pkl* mutant, and biochemical exposure was to either 10-8 M uniconazole-P or no uniconazole-P during seed germination.

Our working hypothesis was that PKL functions during germination to repress genes that promote embryonic identity. In support of such a hypothesis, the transcript levels of two positive regulators of embryogenesis, LEAFY COTYLEDON1 (LEC1) and LEAFY COTYLEDON2 (LEC2) (Lotan, T. et al., 1998; Stone, S.L. et al., 2001), are elevated during germination of *pkl* seedlings (Ogas, J. et al., 1999; Rider, S.D. et al., 2003). Our interest was to find new genes that exhibited PKL-dependent expression, i.e. were up regulated. As such, we had a natural one sided test.

February 2, 2007

第六部分：讀一篇要大改的文章

附上的論文是某作者投稿用的第一稿。他肯投出去，大概以為沒有問題了。我拿掉了原作者的名字和機關，為省空間，也拿掉了所附的圖和表，因為我的目的不是要讓大家搞懂這論文。

這是典型的國人用英文寫的論文——大、小毛病一大堆。最後被接受的形式，和下面所顯示的頗不相同。

請挑出所有你能挑出的毛病（包括大小寫的不當和標點的小錯）。用紅筆在錯的地方註上就好。

The impact of FDI on regional innovation capability: a case in China

Abstract: FDI have been traditionally considered as an important channel in the diffusion of advanced technology. Whether it can promote technology progress for the host country or not is a focused problem in recently decades. This paper analyzes the relationship between FDI and regional innovation capability (RIC). We find that the spillover effects of FDI are not significant as we usually thought. The impact of FDI on RIC is weak; the entry of FDI has no use for enhancing indigenous innovation capability. The research manifests, increasing domestic R&D inputs, strengthening the innovation capabilities and technology absorptency in domestic enterprises are determinants to improve RIC.

Key words: FDI, less developed countries, regional innovation capability, spillover effect, the level of entrepreneurship

1. Introduction

Multinational companies (MNCs) play an important role in the process of global economic integration. Through capital outward, MNCs maintain tight connection with the international economy. The foreign direct investment (FDI) concomitant with MNCs has already become vital source for many less developed countries (LDCs) to obtain international capitals and advanced technology increasingly. Chinese government always attaches lots of importance to attracting foreign investment since the reformation and the "open-door" policy was carried. They implemented the "market for technology" policy, and tried to facilitate the technological progress through attracting inward FDI. China has already

received the total amount of foreign investments over 562.1 billions dollars until the end of 2004, and since 1993 China has been the largest FDI recipient in the LDCs. In 2003, despite of the breakout of the epidemic SARS, the amount of inward FDI reached 53.505 billions dollars which surpassed America and made China become the most inward FDI country all over the world. In 2004, the amount of inward FDI reached 60.63 billions dollars. Now the ratio of FDI to GDP has surpassed 40 percent. [Gong, 2005]

FDI is the important driving force to boost economic development in China. The impact of FDI has penetrated into many aspects of the national economy with the increase of total amount. The negative function appeared gradually such as the homogeneous expansion of Chinese manufacturing and the international trade dissension because of the excessive reliance of FDI and FDI technology. How about the strategic effect of the "market for technology" policy? Does FDI facilitate the technology progress in domestic enterprises? More and more Chinese scholars begin to think about the FDI strategy retrospectively. Many international scholars plunged into the vehement dispute.

As the main way of international capitals fluxion, the impact of FDI on the economy and technology to the host country has been paid more and more attention by international researchers. Many researches showed that FDI was important to facilitate economic development and technological progress for the host country, especially for the developing countries [Kokko, 1994; Kokko et al., 1996; Sjöholm, 1999; Borenztein et al., 1998]. However, the empirical research on the spillover effects of FDI didn't support standpoints above. Recently some empirical researches indicated that the spillover effects of MNCs were weak, and the positive effects of FDI on economy development should have some certain conditions [Young, 1992; Haddad and Harrison, 1993; Kokko, 1994; DeMello, 1996, 1997]. For many LDCs, the correlation between FDI and technological progress or productivity growth was not significant, except for those export-oriented countries [Balasubramanyam, Salisu, and Sapsford, 1996] and those countries having high-level of human capitals [Borenztein, DeGregorio, and Lee, 1998].

Whether FDI can bring the technology progress in China? Chinese scholars also have different opinions. Two scholars in Chinese Academy of Social Sciences Xiaojuan Jiang and Chunfa Wang raised some researches on this problem in 2001 and 2003 respectively. They both investigated the enterprises invested by MNCs and after analyzing the collected data, they obtained contrary conclusions. According to the view of Xiaojuan Jiang, FDI can boost technology progress. There need not any precondition. MNCs will definitely bring their advanced technology, machine and equipments to share Chinese market and consequently enlarge their proportions. While Chunfa Wang considered that the technology spillover effects of FDI were not distinct. Advanced machine and equipments are not equal to technological capabilities. On the contrary, FDI enterprises would decrease and squeeze out R&D activities in domestic enterprises.

Foreign direct investments are carried through by MNCs primarily. MNCs have obvious advantages relatively compared to domestic firms, and hold the most vigorous parts in the world economy. The contribution of MNCs through direct investments on technology transfer is obvious in theory. MNCs combine their own predominance including capital, technology, management skills, marketing channel, R&D and so forth with the advantages in the host country including nature resources, human resource and market scale and so forth to realize the advantage being repaired with each other in the whole world. MNCs

pay more great attention on the R&D of new products. The direct investments of MNCs may bring precious resources including capitals, technology, management skills, R&D capabilities, and the network of international trade for the host country. LDCs attract FDI; bring technology spillover effects through demonstration, imitation, reverse engineering, individual contact, diffusion of management skills, and the exploitation of international market. This is beneficial to shrink the gap in high-technology with developed countries, to upgrade the industrial technology in acceleration, and to fetch up the technology indentations and lag in the course of development. Furthermore, MNCs have stronger technology consciousness and skills; they make use of research institutions, universities and other service organizations more positively, so MNCs can facilitate the construction of national innovation system for the host country to some extent.

FDI may also bring negative spillover effects. The technology spillover effects of MNCs are likely to very feeble. It is one kind of crucial means that MNCs invest in LDCs in order to keep their own technology advantages and at the same time to share the cost advantages with LDCs. MNCs mostly invested the capital-intensive and technology-intensive industries which were laggard in China under the consideration of the global strategy and they aimed to occupy Chinese market in the long term. MNCs have obvious competitive advantages compared to domestic firms. They always bring more advanced technology and equipments through FDI. However, the spillover effects can't come into being automatically. Imbriani and Reganati (1997) represented that the spillover effects from FDI enterprises were in inverse proportion to the size of technology gap between foreign and domestic firms from the Italian evidence. Kokko (1994), Kokko, Tansini and Zejan (1996) also found that a positive and statistically significant spillover effect only in plants with a moderate technology gap. There almost were not any spillover effects when the technology in foreign firms was much higher than domestic firms according to the research of Mexico and Uruguay. Generally speaking the introduced technology must be suitable with the factor endowments in the host country. If the technology gap were too large, although there may be many opportunities for domestic firms to imitate and learn, domestic firms may have not enough technological capabilities to absorb and imitate advanced technology from FDI enterprises and can't make the advanced technology be endogenous. Thus it results in little spillover effects. Because of the stickiness of information, most technology and knowledge are tacit knowledge. Only through practice can they be mastered. Just as the study of Borensztein (1998) showed that the introduction of more advanced technology and the requirement of absorptive capability in the host country were twinborn factors of economic growth. FDI was more productive than domestic investment only when the host country had a minimum threshold stock of human capital. DeMello (1999) discovered that FDI had positive effects to the countries with high technology and had negative effects on the followers. So the contribution of FDI not only depends on the technology level brought by foreign investment, but also depends to a large extent on the ability of absorption and assimilation of advanced technology of domestic firms. To acquire the spillover effects of FDI, the host country must be in possession of enough human capital that has received good education and trainings, and proceeds to reform and innovate constantly after utilizing the acquired technology effectively.

Because MNCs possess more advanced technology and management skills compared with domestic firms, their entries monopolize the original competition market in the host country, squeeze out the domestic firms in the industry and lower the market share of domestic enterprises. Kokko (1994) found that no evidence would show the positive effects of FDI on domestic productivity growth when the market share of MNCs was big and the

technology gap was too large. Demonstration and competition effects are really effective mechanism of spillover effects, but this must be based on corresponding technology capabilities in domestic firms. MNCs invest, incorporate and purchase the potential rivals in the host country, which may cause the innovation activities in the purchased enterprise be decreased, be transferred or be closed; furthermore decrease the regional R&D activities, make us excessively rely on foreign countries about key technology and bring the negative influence.

The majorities of MNCs invest in our country and engage in only the production and operation of the final product at present. Those key intermediate products were usually supplied by MNCs internally. So the linkage effects were limited. MNCs integrate with the Chinese base of manufacture very slowly to protect their own technology and the manufacturing expertise. Also, they seldom have relation with Chinese universities and research institutes. MNCs usually take the global production strategy and arrange every kind of functional behaviors in the whole world to excavate the competitive advantages on each step. Chinese enterprises are just one of its manufactured chains in the global production. This is manufacture transfer but not technology transfer. Technology circulation happened only within the MNCs. This is one kind of floating economic with bad ground-work. The advanced technology and equipment can't be transferred to domestic firms automatically. The technology stream can't be transformed into endogenous technology capabilities easily.

The purpose of this paper is to analyze quantitatively the relationship between FDI and regional innovation capabilities (RIC) and to find the determinants of RIC. We will empirically do research on the correlation between FDI and RIC using data of each province in China; verify whether the more of inward FDI in a province will lead to a higher level of innovation capability. The paper is organized as follows. Section 2 presents our research method. Section 3 gives the concept of regional innovation capabilities. Section 4 researches empirically the relationship between FDI and RIC using multivariate statistics analysis. Section 5 constructs empirical model to analyze the impact of FDI on RIC further. Section 6 analyzes the effect of FDI to the level of entrepreneurship. Section 7 concludes.

2. Research design

In this paper, we focus on the relationship between FDI and RIC and highlight the role of FDI in the RIC in China. What is the regional innovation capability? This is the principal issue we should conduct on. So we make certain the concept of RIC firstly. In fact, the connotation of RIC is very abundant. Inward FDI is just one of the main factors of RIC; it means the ability of technology introducing and transfer. On the basis of the comprehension on the meaning of the concept, we establish a series of indicators to reflect RIC completely and objectively. Then we construct synthetic evaluating function. The score of the constructed function can represent RIC accurately.

We use principal component analysis of multivariate statistics to obtain synthetic evaluating function. We adopt the first principal component (PCR1) as the synthetic evaluating indicator to appraise RIC and calculate the correlation between FDI and RIC. Because PCR1 is the weighted sum of each original variable, and its variance is the biggest, it can

reflect majority of the information of all original data. The factor loadings of each original variable on PCRI are the correlation of each original variable and PCRI. Through this analysis we can also obtain the determinants of RIC. We will elaborate on the characteristics of some typical regions. We conduct on detailed compare and analysis to summarize the rules of development of RIC.

We establish the empirical model with regression analysis to compare the importance of each factor that be related to RIC using the number of patent applications to measure RIC. Then we use correlation analysis through scatter diagram and the determinate coefficients of regression model to manifest the impact of FDI on the level of entrepreneurship. All analyses are dealt with SAS statistical software.

All data used in this research are obtained from China Statistical Yearbook, China High-tech Industry Statistical Yearbook, and China Science and Technology Statistical Yearbook. These three yearbooks are compiled by National Bureau of Statistic of China. Some of the variables can't be directly obtained and we proceed to some simple computation. All data are the most up-to-date currently available.

3. What is the regional innovation capability?

Technological innovation is a concept of economic category; it means the economic-technological activities including R&D, production and commercial applications of new technology (include new product and new craftwork). Technological innovation is one kind of commercial activities which creates new economic value by means of new technology (entirely new or modification). It is the first commercial application of new technology, and realizes the combination of economy and technology. Technological innovation has three particular characteristics. First, it emphasizes that the degree of market realization and acquisition of the business benefits are the ultimate standard to verify whether the innovation is successful or not. Second, it emphasizes that it is a systematic engineering from the research and development of new technology to the first commercial application. Third, it emphasizes that the business enterprises are the subjects of technological innovations.

Technological innovation has two types of realization modes, namely the independent innovation mode based on R&D, and the second-innovation mode based on introducing in and assimilating advanced technologies. The investments on independent innovation of new technology and new products are considerable, and the risk is also stupendous. It needs very strong capability of R&D and needs large amounts of funds for support. Because of the large technology gap between developed countries and China, the second-innovation mode is usually taken in China. With the fast development in technology and science, the scope of technological innovations become more and more wide, the competence of making use of the exterior knowledge has become an important part of technological innovation capability.

The regional innovation vitalities are enslaved to innovation capabilities. The regional innovation capabilities are the potentialities of producing streams of innovations related to commerce in a region. They refer to the capabilities of converting knowledge into new product, new craftwork, and new service. The regional innovation capabilities are not

merely the capabilities of science and technology, nor just the technological competition ability, for they pay more attention to the economic applications of new technology. The strong technological competition ability of a region doesn't mean that the innovation capabilities are strong too. The regional innovation capabilities are made up of those factors as follows: Technology human resource, the creators of knowledge and employees who grasp the subjects of craftsmanship; the ability of knowledge fluxion, that is the ability of making use of all kinds of resources in the world constantly, and the ability of knowledge fluxion among each innovation units; the capabilities of technological innovation in enterprises; innovation environments and the economic performances of innovations, that is the output of innovations.

The innovation capabilities will decide the long-term economic competitiveness in a region. The innovation capabilities are the most important factor to explain the differences of the degree of economic prosperity among high-income nations. For LDCs to say, obtaining and developing the technology has become an essential driving force to improve the competitiveness [The Global Competitiveness Report 2002-2003]. For LDCs to say, introducing in advanced technology and assimilating, realizing the second-innovation is the shortcut to improve the regional innovation capabilities and international competitiveness.

4. The correlation analysis between FDI and RIC

4.1. To establish the series of indictors of RIC

To evaluate the correlation between FDI and RIC accurately, we need to describe RIC exactly first, so we must establish a series of indictors to evaluate RIC. The establishment of the series of indictors of RIC is the further step of comprehension on the essence and the meaning of the concept of innovation, and is also the development of innovation theory. The connotation of RIC is very abundant; the determinants include education, science and technology resources, innovation capability in enterprises, regional synthetic strength and information condition; and also include the regional policy and management skill. The series of indictors of RIC must reflect the present conditions and the utilization efficiency of regional knowledge and technology objectively and try to evaluate the RIC completely and objectively. The indictors being chosen should be brief and terse. On the basis of reflecting RIC accurately, try to select those synthetic indictors which have the commonness and are maneuverable. Therefore, we select 22 indictors altogether to measure RIC as listed in table 1.

Table 1 about here

4.2 The principle component analysis

We use the principal component analysis to establish the synthetic variable evaluating RIC of each province. Its excellence is that the weights are based on the inherent configura-

tion relationships of the variables from the data analysis, there doesn't have any influence of subjectivity, and is objective perfectly. This is beneficial to analysis and appraise synthetically. We analyzed the data of 31 provinces in China about above 22 variables with principal component analysis, using SAS software to deal with. The biggest eigenvalue of the correlation matrix values 11.1167 and the first principal component reflects 50.53

Table 2 about here

By the eigenvector of the biggest eigenvalue in table2, we can get the expression of the first principal component (PCR1) as following:

$$\begin{aligned} PCR1 = & 0.1237RFUND + 0.1338FDI + 0.2793TVMARKET - 0.0814EXPPT \\ & + 0.1971EXPIT + 0.1107RR\&D + 0.0818RS\&T + 0.2875UTILITY \\ & + 0.2380DESIGN + 0.0727OVEQUIP - 0.0136EXPTR \\ & + 0.2426NEWSALES + 0.2494NEWRATE + 0.2589COLLEGE \\ & + 0.2781HIGHEDU + 0.2611BOOKEXP + 0.2685NS\&E \\ & + 0.2237NPTE + 0.2303NHTE + 0.2597NNE \\ & + 0.2871PGDP + 0.1447LPRODUCT \end{aligned}$$

The expression shows, PCR1 is the weighted sum of each original variable, most of the coefficients are positive excluding two indicators of EXPPT and EXPTR. Indicator EXPPT means the average expenditure on purchase of domestic technology of each enterprise. Generally speaking, the higher of technological capabilities, the less domestic technology enterprises would purchase. So the negative value of the coefficient is reasonable. Indicator EXPTR means the average expenditure on technology renovation of each enterprise. The coefficient is negative, but the number values only -0.0136, so its influences is not big. We can neglect it. Thus PCR1 can reflect the innovation capability of each province synthetically. So we use PCR1 as the synthetic evaluating indicator to appraise RIC.

4.3 The impact of FDI on RIC

The loadings in table 2 are the factor loadings of each original variable on PCR1. Factor loadings are the correlations of each original variable and the principal component in fact. The correlation coefficient between FDI and PCR1 is only 0.4462, there has no significant correlation between FDI and RIC. By table 2 we find that the variables having high relationship with RIC include: TVMARKET (respects the ability of technology transfer), UTILITY and DESIGN (respect the ability of design in enterprises), NEWSALES and NEWRATE (respect the ability of innovation output in enterprises), COLLEGE, HIGHEDU, BOOKEXP (respect the caliber of employees), NS&E (respects the human resource in S&T), NNE (respect the level of entrepreneurship) and PGDP (respects the performance of innovation). Therefore, for a nation or a region to say, the determinants of innovation capability include: paying attention to the investments of S&T, the ability of upgrading the human resource in S&T and making full use of all kinds of resources

of S&T; the technological innovation capability in enterprises; having good environments beneficial to innovations, the mechanism and system that is open and can make full use of the local special resources and the global S&T resources; the outstanding performance of local economic, which become the powerful pull force of innovation. We analyze the correlation between FDI and RIC further using the synthetic variable obtained from the principal component analysis, drawing the scatter diagram, establishing the regression model and obtaining the regression curve. See as figure 1.

Figure 1 about here

We can observe the relationship between FDI and regional innovation capability clearly by picture 1. From the scatter of points that represent each region in picture 1, we can discover that there have many points far away with the regression curve, such as Beijing, Shanghai, Tianjin, Guangdong and Jiangsu. The determinate coefficient R^2 is only 0.1991, it indicates that FDI can explain 19.91

By picture 1, we can also find that there have a lot of points almost gathered together. These points represent the provinces in which both FDI and RIC are all lagged behind. These provinces locate in the centered or western region of China, are less developed in economy because of the lagged notion, policy and geographic location. The behindhand innovation capabilities are mainly induced by the lagged economic conditions. To eliminate the influence of economic development, we calculate the correlation again canceling the regions in which both FDI and RIC are all lagged behind and obtain the correlation is minus 0.3626. This indicates that FDI has the negative impact on RIC.

FDI in Guangdong and Jiangsu are the biggest. The amount of FDI in Guangdong is 11.334 billions dollars in 2002, and 10.19 billions dollars in Jiangsu. The amount preponderated over other regions greatly. But their regional innovation capabilities lie in the fourth and the fifth respectively, and just bigger than the latter regions a little. Then we compare each original variable carefully and discover that on three variables that respect the caliber of employees, which are number of college and higher level per 10000 populations (COLLEGE), number of graduates from institutions of higher education per 10000 populations (HIGHEdu), the books sales per capita (BOOKEXP), Guangdong located in 12 and 13 and 21 respectively in the whole country. Variable that respects the human resource in S&T, number of scientists & engineers per 10000 populations (NS&E), Guangdong located in No.7. And Jiangsu located in 20, 6, 4 and 5 on above variables respectively. These show that the level of human capital doesn't match the amount of FDI in Guangdong and Jiangsu. Whether the stock of human capital is high or low will decide the spillover effects directly, and influence a nation's independent innovation capabilities and potentiality directly. Only when the volume of FDI match the stock of human capital and technological capabilities, can the RIC be developed and enhanced.

The region with the strongest RIC is Shanghai, followed by Beijing, while FDI in these two regions listed in the fourth and the eighth respectively. What drive the innovation capability in Beijing and Shanghai? Obviously we can't get the reasonable explain only from the FDI. The superiority in Beijing include the abundant S&T resources, strong abil-

ity of creating new knowledge, attaching importance to the investments for S&T, having the most excellent employees, and the good environments for innovation. While in Shanghai, there have a good foundation for industrial innovation; the ability of technological innovation of enterprises in Shanghai keep ahead in the whole country; and there have the strong financial strength and powerful capital advantage. Zhejiang is the province in which private economy developed very well, the capability of innovation in private enterprise is extraordinary. Zhejiang developed quickly because of the dual superiorities of S&T and the system. People have great enthusiasm to start-up an enterprise. There have the perfect system for regional scientific and technological innovation, investment and financing and supporting for the talented person in Zhejiang. Zhejiang is the apotheosis of making full use of local resources to improve the innovation capability.

Therefore, being open to external, attracting FDI is not the unique means to increase RIC, and it isn't the important means either. The speed of reformation and the degree of marketization in economy will decide the level of innovation, the notion and system will influence the innovation capability. The determinants of the innovation capability include the original drive force of innovation, the technological innovation capability in enterprises, the ability of making full use of the local special resources and all kinds of S&T resources, having the good environments beneficial to innovation, perfect mechanism and system for regional scientific and technological innovation. Besides above factors, the local economic condition is the strong pull force for innovation. On the other hand, the level of GDP will be high in the regions that have strong innovation capability, and the international competitiveness will be strong too.

5. The empirical model of the effect of FDI to the innovation capability

We measure the innovation capability with various different variables in the above analysis. But scholars often use the patent data to measure the technological innovation, because when an inventor or creator apply patent to the patent censorship, this is usually the potential sign of economic value of innovation and representation of the innovation capability. Furthermore the patent data are complete, accurate and can be obtained easily.

In less than 20 years, China has made tremendous progress in establishing a legal system for the protection of innovation. China's first patent law was enacted in 1984 and came into effect in 1985. Since then, the law has been amended twice. Since the passage of the 1984 patent law, the central government has issued over 20 regulations and guidelines so as to promote innovation activity in China. Today's patent law in China is pretty much in line with the international standard [Kui-yin Cheung, 2004].

The patent law of China divides patents into three categories: invention, utility model, and external design. The invention patent refers to the new techno-project put forward on the product, method or its modification, which can form the products having independent intelligent property. The utility model patent means the new practical techno-project put forward on the shape, structure of the products or their combinations. The external design patent is the new design which is full of pleasant impression and suited for the application in industry about the shape, pattern, color, or their combinations of the products. The innovative level is relative low of the utility model patent and the external design patent. The term of protection is 20 years for invention patents and 10 years for utility model and

external design patents. Among three types of patents the invention patents are regarded as major innovations, they can represent the innovation level mostly.

The number of patent authorized was influenced by the abilities of the patent censorship, so in this text we use the number of patent applications in our country to measure the regional innovation capabilities.

We use the following model to estimate the spillover effects of FDI on innovation capabilities in China.

$$Patent = f(FDI, PGDP, S\&TEXP)$$

We use the number of patent applications (Patent) as a measure of innovation capability. FDI refers to the realized values of FDI lagged one period considering that FDI inflow to China impacts on domestic innovations within a short period of time. As measures of input to R&D activity, we use expenditures on science and technology development (S&TEXP). Finally, considering the fact that different provinces are at different stage of economic development so that their innovation capabilities should also differ, we include the level of per capita GDP (PGDP) in our estimation.

The model can also be expressed as: $Patent = A \cdot FDI^\alpha PGDP^\beta S\&TEXP^\gamma$, then we take logarithm in both sides, and the model can be written as following:

$$\ln Patent = C + \alpha \ln FDI + \beta \ln PGDP + \gamma \ln S\&TEXP + u$$

Where C is the estimated intercept of the equation, u is the statistic error, the coefficient α, β and γ is the elasticity of the increase of FDI, PGDP, S&TEXP to the patent increase respectively. For convenience, the equation was written as follows:

$$Patent = \beta_0 + \beta_1 FDI + \beta_2 PGDP + \beta_3 S\&TEXP$$

The coefficients $\beta_1, \beta_2, \beta_3$ measure magnitude of the influence of FDI, PGDP and S&TEXP respectively. The data are taken from China Statistical Yearbook and China Technology Statistical Yearbook, covering 30 provinces. Tibet is excluded in our analysis because most of the relevant data for it is either not available or zero during the time period examined.

We used SAS software to estimate the equation with the ordinary least squares (OLS). And we estimated the equation for each type of patent (invention, utility model, and design) as well as the total patent applications.

Table 3 about here

By table3, the determinate coefficients R^2 of four models are all above 0.77, and the values of F-statistic are significant at the level of 0.0001. This shows that four models all have great statistic significance and they can explain the variations of all types of patent

applications more than 77 percents. For the amount of invention patent, the coefficient of FDI is positive, but it is statistically insignificant even under the level of 0.10. So FDI has no significant effect to the amount of invention patent. There has positive effect for FDI to the amount of utility patent, and has significant effect to the amount of external design patent and total patent. S&TEXP has significant effect to the amount of all three types of patent and the total patent under the significant level of 0.0001, except the external design patent is under the significant level of 0.05.

The results manifest that S&TEXP is the most important factor to increase the innovation capabilities. FDI has positive spillover effects to some extent, but has no significant effect to increase creative inventions and the inventions having independent intelligent property rights.

6. The effect to the level of entrepreneurship

We use the number of new registered enterprises (NNE) and private technology enterprises (NPTE) and high technology enterprises (NHTE) as a measure of entrepreneurial level. FDI refers to the realized values of FDI lagged one period considering that FDI inflow to China impacts on innovation activities within a short period of time.

The effect of FDI to the entrepreneurial level can be showed as the following figures.

Figures 2, 3, 4 about here

We observe three scatter plots in which each point represent each province, and find that there have many points far away with the red regression curve. Three correlation coefficients between FDI and the number of all kinds of new enterprises and the determinate coefficients of regression equations are all very low. So we can make conclusions that FDI has no direct significant effect to the level of entrepreneurship. The more FDI will not bring the higher entrepreneurial level.

This result may be induced by the crowing-out effect of FDI. Because of the high risk to start-up an enterprise, a successful entrepreneur must be adventurous pioneer daring to take the tremendous risk and failure. The presence of foreign enterprises may provide higher wage for employees and increase the competition in the market, this increase the difficulty of successfully starting an enterprise. So it is likely that many people having certain entrepreneur spirits would rather select the secure and steady job in foreign enterprises than undertake the risk to carve out.

7. Conclusions and policy implications

As far as LDCs are concerned, introducing in advanced technology is the shortcut to facilitate the technology progress. FDI has been regarded as an important channel for technology diffusion. However, the results in this paper demonstrate as follows.

First, the correlation between FDI and RIC is insignificant statistically. The impact of FDI on the regional innovation capability is inappreciable. The regions which attract more FDI haven't the higher regional innovation capability. Only when the volume of FDI match the stock of human capital and technological capabilities, can the RIC be developed and enhanced. This finding is the same to Borensztein (1998). Attracting FDI is not the unique means to increase RIC, and it isn't the important means either. The determinants of RIC include the original driving force of innovation, the technological innovation capability in enterprises, the ability of making full use of the local special resources and all kinds of S&T resources, having the good environments beneficial to innovation, and the local economic conditions.

Second, the results of the statistical model indicate that investment in R&D activities is the most important factor to enhance the innovation capabilities. FDI has positive spillover effects to some extent, but has no significant effect to increase creative inventions and indigenous innovation capabilities.

Third, through three scatter diagrams and correlation coefficients between FDI and the number of new enterprises and the determinate coefficients of regression equations, we find that FDI has no direct significant effect to the level of entrepreneurship.

To sum up above results, MNCs have the technological advantages to some certain; the inward FDI will facilitate the technology progress and improve the regional innovation capabilities in some degree. But we can't think blindly that we introduce in FDI the more the better. The advanced technology and equipments of FDI can't be transferred to domestic firms automatically. There have no significant correlation between FDI and RIC, the more of FDI will not necessarily bring the stronger innovation capabilities and spirits. So we can barely rely on FDI to improve RIC. The determinants to improve RIC include such factors as following: increasing domestic R&D investments, enhancing the stock of human capitals, improving the technological innovation capabilities and technology absorbency in domestic enterprises, having good environments for innovations.

The findings in this paper may provide some insights for both the host countries and foreign investors. Based on above analysis we put forward some policy implications as follows, which may also be effective to other LDCs. Firstly, it is urgent for China to improve the stock of human capital and indigenous R&D capabilities. Whether the stock of human capital is high or low will decide the spillover effects directly, and influence a nation's indigenous innovation capabilities and potentiality directly. It is necessary for China to enlarge investments in fostering human capitals, especially the investments for fostering R&D personnels. It is necessary for China to strengthen cultivating the independent R&D capabilities, particularly supporting for the R&D activities of techno-intensive industries. These are important to shrink the technology gap between domestic enterprises and foreign capital enterprises. Only when the host country has certain technological talented persons, can MNCs arrange R&D projects in the host country and train the native high technological talented persons, to reduce the cost of human resource and make products locally better. Therefore making native innovative systems perfect, cultivating R&D capabilities of domestic enterprises positively, promoting and stimulating independent innovations in domestic enterprises are the basis for improving technological progress from making use of the spillover effects of FDI. Secondly, Chinese government should focus on the quality of inward FDI and insist on the sticking point to advance independent innovation capa-

bilities. Try to urge and guide FDI to be on the trajectory which is beneficial to improve our independent innovation capabilities. Thirdly, China should try to create a fair competition environment for domestic enterprises to compete with foreign enterprises, and keep sufficient rivalrousness of the market. Under the intense competition environments it may force MNCs to increase the degree of technology transfer to our country.

References are omitted, end of paper for correction.

September 4, 2006

第七部分：讀一篇關於自助法的論文

在這一部分，我們選出 B. Efron 的一篇在 1979 年的論文來細讀——主要在探討他的寫法，論文寫得好的境界有三：初級為易售，讓評審覺得可以刊登；再進一步是真有內容，此時，即使寫得略差，也不會受太大的影響。最糟是其實沒有實質內容但寫得花團錦簇¹。

此文為 1977 年的 Rietz 的獎座論文。這是國際數理統計學會 (Institute of Mathematical Statistics, IMS) 在它的年會中所邀請的主要演講。這演講會在會場的最大廳內舉行，且同時沒有其他的論文在宣讀。能受到邀請的學者皆非等閑之輩——事實上此文在 *Annals* 上根本不會被拒絕，因為否則 IMS 豈不是自己打自己的嘴巴？

Rietz 講座因為要面對較大的聽眾 (雖然 IMS 的會員都是以數理統計為主，但也不是人人皆是高手)，因此這樣的論文也不會寫得太深。但更不會太淺，因為否則又豈不弱了被邀演講者的名頭？對這類的文章，作者有較大的期望，希望能 (對統計科學) 發生影響。因此他會寫得不僅能懂而且能用，且同時讓讀者覺得這文章有學問。

這篇文章的全文，見 *Annals of Statistics* 7, 1-26。下面的註腳，要配合原文來讀，可以當作此文的「導讀²」來用。我們的目的不是教你自助法，而是介紹你這論文的寫作。

論文的標題是 “Bootstrap methods: Another look at the jackknife”。用了兩個名詞：bootstrap 和 jackknife。這兩件事都是「重抽法 (resampling method)」。在演講的時候，jackknife 已廣為人知，但 bootstrap 卻是新的東西。以下為我的註腳。

p.1, abstract, line 1³. random sample 指的是 $X_1, X_2, \dots, X_n \sim \text{iid } F$ 。此處直接引入 $R(\mathbf{X}, F)$ 這個統計量，這個統計量是本文今後要討論的主旋律。注意到它有未知的分布 F 在內。這和我們熟知的統計量如 $\bar{X}, s^2$ 有異——它們不包含未知參數⁴。但目前你可將它想像為 $\bar{X} - E(X)$ 或者 $s^2/\text{Var}(X)$ 之類的統計量。

p.1, abstract, line 4. 在此直接提 jackknife，並和本文目的做一個區隔。本文可看成為 jackknife 的推廣，但作者並不這樣看，這樣看就把本文看小了。他將 bootstrap 看成為一種自然的功法，而將 jackknife 看成為一種類似的，但較低層次 (因此反而較複雜、較不清楚明白) 的工具。當然，在寫此文的時候，多數人知道 jackknife，而 bootstrap 則正要推出，知者甚少⁵。所以他用 jackknife 作為比較也是合理的。此時，他對 jackknife 既不能捧也不能罵，因此說 “is a linear approximation of jackknife”——但此話可不能空口說，後文得要有本事證明。

p.1, introduction, line 5. 用 “attempt to explain” 是恰當的用語。這回答了「何謂

¹「花團錦簇」是舊小說中形容八股文寫得好的用語。

²這裡當然充滿了我的看法 (如果不是偏見)，也不盡然是作者的原意。此文我每讀一次，都有些新的領悟，因此這一次不到之處，必然也是有的。

³這指從 abstract 起第一行，下同。

⁴在無母數問題中， F 可視為未知參數，但它是「無窮維 (infinite dimension)」的。

⁵事實上真在這一行中的人應該早已知道。開玩笑，Stanford 的統計系正預備推出的一套方法，如果你還想在這個行業競爭，豈有沒聽過一點風聲之理？在第一線作戰的戰士，耳朵當然是豎起來的。

jackknife」這樣的較深層次的問題。

p.1, introduction, line 8. “more dependable” 在後面可是要用例子說明的。故下一句立刻說：對於 sample median 而言, bootstrap 就比 jackknife 有用——因為已知 jackknife “is known to fail”。

p.1, introduction, line 10. 立刻又說, 對於 linear discrimination 中, bootstrap 比 cross validation (另一種已知的重抽法, 常在 linear discrimination 中使用) 要好。

p.1, introduction, line 13. 此一段又提 jackknife, 說可以用 delta method 來看出 jackknife 可以視作為 bootstrap 的一次微分。並說因為可以這樣看, 不止對 jackknife 可以提出理論基礎, 並因此可以回答一些原先 jackknife 需要回答但一直說不清楚的問題。此段的末尾還說, 若有額外條件——如已知 F 為對稱或者 smooth——該如何做 bootstrap。最後, 更進一步提到, 對於迴歸, bootstrap 又該如何做⁶?

p.1, introduction, line 21. 此一段說本文以舉例為主, 對於一般性的 bootstrap 理論著墨不多⁷。

p.2, line 3. 此一段承認以前各作者的功勞。如 [5, 6] 為 “ideas closely related”。注意他不承認 100% 相同, 因此否則自己就沒有貢獻了。又說 [13] 中的方法可以導出本文的公式 (3.4), 又說 [10] 中的 infinitesimal jackknife 直接和本文的 method 3 相關。這一段是預告, 將後文的 method $i, i = 1, 2, 3$ 都一一交待。

注意到 [10] 並非一般論文, 而是 Bell Labs 的 technical report。一般在重要論文中是不引用技術報告的, 但 [10] 卻是有名的一篇文章, 內行人都知道。所以可以在本文中大方引用。

本文最後提出 parametric bootstrap, 以印證本文所討論的 non-parametric bootstrap 為主, 但 parametric bootstrap 的道理是類似的, 且和 R. A. Fisher 的 information bound 一致。注意到 Fisher 基本上已解決了所有傳統型的 parametric inference 的問題, 可由此顯示出 bootstrap 的不同凡響——可以大小古今通吃。

p.2, section 2, line 6. 此處 “plays no role” 一句, 是為了後面的例子 (如相關係數) 做伏筆。在此也等於是說, 此例雖然 F 為一維, 但 bootstrap 的應用則不限於此。

p.2, section 2, line 12. 此一段用簡單的一段文字介紹 jackknife, 並暗示 jackknife 的不足: (1) 只有興趣在 bias 和 variance; (2) jackknife 只有興趣處於 (2.3) 或 (2.4) 這種形式的統計量⁸。

p.3, lines 3-16. 這一段直接介紹 bootstrap 的運作方式 (但 line 10 中所提的 “drawing samples of size $n - 1$ without replacement” 是一直沒有講清楚的)。

p.3, lines 17-23. 這幾行不那樣易懂, 但卻是 bootstrap 的基本道理。“Fisher consistency” 並不那樣 well-known, 但打出 Fisher 的大名, 不懂的人只好承認自己的學問不

⁶這裡提一下, 一般的統計方法, 若無法推展到迴歸型的數據, 就不能算是完整。所以他提「能做迴歸」, 是有「本研究基本已完備」的暗示。

⁷不要以為 Stanford 的人不做 bootstrap 的理論。相信此時作者必有些學生在努力。甚或已有一些成果。要比這些人快, 你需要更快, 並且要消息靈通。去國外某學術單位訪問, 這是一個目的。

⁸其實本文中 bootstrap 也多在對付這類統計量, 但因在架構上 $R(X, F)$ 是廣泛得多, 所以看起來較高級。

夠。bootstrap 的精神是：用 R^* 的分布來近似 R 的分布。為何這竟是對的？我曾有較數學的表示法如下。 $R(\mathbf{X}, F)$ 的分布可以用它的 characteristic function (ch.f)

$$Ee^{itR} = \int e^{itR(\mathbf{x}, F)} dF(\mathbf{x}) \quad (1)$$

來表現。同理， $R^* = R(\mathbf{X}^*, \hat{F})$ 的分布，則用類似的 ch.f

$$Ee^{itR^*} = \int e^{itR(\mathbf{x}^*, \hat{F})} d\hat{F}(\mathbf{x}^*) \quad (2)$$

來看。在 (1) 和 (2) 之間，有兩點相異： $\mathbf{x}$ vs. $\mathbf{x}^*$ ， F vs. $\hat{F}$ 。但仔細看一下，(2) 中的 $\mathbf{x}^*$ ，其實是一個用來指示對誰積分的 dummy variable。若在 (2) 式裡的積分中令 $\mathbf{x} = \mathbf{x}^*$ ，則 (2) 變成

$$\int e^{itR(\mathbf{x}^*, \hat{F})} d\hat{F}(\mathbf{x}^*) = \int e^{itR(\mathbf{x}, \hat{F})} d\hat{F}(\mathbf{x}). \quad (3)$$

比較 (1) 和 (3) 則知道若是 $F = \hat{F}$ ，則 (1) 和 (3) 是完全一樣的。因此二者的分布在此時是一樣的。

p.3, lines 24-29. 這一段的好句子是 “let the varying degree of success of the examples speak for themselves” —— 這又暗示了本文大量舉例的原因。

p.3, line 30 及以下。這是做研究的基本方法之一：用最簡單的 non-trivial case 來驗證理論。此處 $X_i \sim \text{iid } b(1; p)$ 是幾乎不能再簡單的情形，用前述的 bootstrap 方式來做，是否會得到理想的結論？如果有，可以向下面繼續做；如果沒有，就得去找原因。

p.4, line 4. 用 “universally familiar” 一語帶過多少教本上的結果，

p.4, lines 5-16. 這一段是可以硬算的。此處舉的是一個較複雜的例子。注意到因為 $\bar{X}$ 的情形等於已做過，雖然那是最簡單的情形，此就再做就沒有意思了。因此此處用 second moment 未討論，以示和 binomial case 有基本的不同。至於引用 [3]，也是寫作技巧之一。這 Cramer (1946) 可是經典的古典教本。你若是隨便引用一本 Statistics in Business，看在評審人的眼裡，你的份量就全不是那回事了。

p.4, line 19 及以後。此處提出三種方法來算 R^* 的分布。最後當然大多數人會用 method 2，但方法一是用來印證用的：能做解析計算時還是要做解析計算。Simulation 終究不如 exact 的計算結果。方法三其實是無用的，在本文中有用，是因為我們要靠它建立 bootstrap 和 jackknife 間的關係。

但此文若沒有和 jackknife 的一段姻緣，文章的支撐就少一些。所以這一文字是可加上的。此外，用三個角度來看 bootstrap，也表現作者上功夫。

p.4, line 29 及以後。介紹本文的其它內容。主要指的是 2-sample 問題的處理。雖然指的是 section 4 和 section 6 (沒有連貫)，但二者都是 2-sample problems，故可放在同一段落。

p.5, section 3. 本節以硬算為主，要點在 p.6, lines 8-13. 這幾行用明確的事實說明 “bootstrap works for sample median, while jackknife does not”。注意到 [11] 也是經典教本 (且引用時直指 p.237 —— 這是應該的。直接引用某書的時候，應結出較細的資訊。如 [3], Section 27.4 (p.4, line 11).)。這表示兩件事：(1) 作者對讀者負責，不希望你在

一本 500pp 的書中去找他所引用的某件事, (2) 表示作者真的讀過, 不是鬼混的。至於指 [14], p.8 有錯一事, 一般是有風險的。但若你真的挑得出錯, 也可以明說。此處 [14] 的作者 Miller 是 Efron 的博導, 二人關係良好, 是無所謂的。

p.6, line 14 – p.8 end of Table 1. 這兩頁若是由一個小教授來寫, 多半會被要求刪去。但 Efron 是有名的大教授, 所以可以多用篇幅。重點在 table 1 後面所說的 “the most notable feature ...”。有這一句才能 justify 這兩頁的篇幅。並同時強調了最簡單的 bootstrap, 其實非常不錯。至於極力地將問題中的額外條件 (如對稱、平滑) 用進 bootstrap 的努力, 其實是沒有甚麼大道理的。

如果在 1979 年, 如果你數學不錯, 一個立刻可以做的題目是: 用數學來證明或反證這兩件事。此文只用了 $n = 13$ 而已。那麼, 當 $n \rightarrow \infty$ 時, 情形會如何? 對數學好的學者, 這不會太難。而在 1979, 這類文章多半可以刊載。我的直覺是: (1) 你必需手脚要快, (2) 二者的差別應在 $O(n^{-1/2})$ 的項以外, (3) 這種文章即使能做也沒甚麼了不起, 你只是證明你技術不錯。

但現在是 2006, 你必須好好查一下才會知道此事有沒有已經發表。而 p.8 的最後一段是作者的感覺, 雖然這不會是無的放矢, 但也不會有真正的大意義。明顯地, Efron 目前根本不想追究這些現象。

這一節只有兩個要點: (1) 對 sample median 而言 (因為已知標準答案), bootstrap 較 jackknife 要好: 前者做對, 後者做錯。(2) 簡單的 bootstrap 不比複雜的 bootstrap 要差。這裡面, (1) 是數學證明, (2) 則只是模擬的結果。但作者巧妙地安排 bootstrap 的出場場景, 並用 jackknife 來陪葬, 雖然不是自己的計算 (作者大方承認, 本節的主要技術來自 [13])。這是重要的, 不要掠美 —— 這是學術論文的規矩。

p.9, section 4.

本節又舉了一個例子: 對於 discriminant analysis 中的 error rate 的估計量, bootstrap 要比 cross validation 要強。

全節都是在用 simulation 來做 (道理當然是算不出解析的結果)。這個例子和前面的 sample median 不同, 否則就沒有甚麼意義。它的不同點在: (1) 這是 2-sample 問題; (2) cross validation 和 jackknife 是不同的事; (3) 我們在此算不出 exact 的結果。因此 method 2 是唯一的方法。並且, 到目前為止, method 2 尚沒有舉例。

主要的結論在 p.10 的 Table 2 中: 差別是 0.09 vs. 0.03, 有三倍的不同。

本節沒有理論, 但模擬的步驟交待得十分清楚 —— 我可以根據作者的描述, 寫出一個程式同樣做出作者的結果。這也是科學論文要點之一: 所描述的實驗⁹ 資料, 應豐富到別人也能據此做出同樣的實驗的程度¹⁰。這是「可重覆性 (repeatability)」的要求。

此節除了選題正確 (需要一個 well-known 的問題, 它還得有一個 well-known 的解, 這個解還得比不上 bootstrap), 且本例的各種條件都尚未被討論過, 因此除了寫得仔細之外, 並不算難。可提出的是, 在 p.10 的 (4.7) 中, 要用 $[X_i \in B^*]$ 而非 $[X_i \in B]$; 在 p.11, line 19 處, 要指出該用 m 而非 $m - 1$ 。這些都是應該有理由的, 想來真正的專家都知道。但我們只講文章的寫法, 而非導出結果的數學。我們就不提了。

⁹ 模擬當然是一種實驗, 儀器是電腦。

¹⁰ 道理是: 別人有辦法查證你的實驗是否做對了。

要注意的是 p.9, (4.1) 中的「符號成對」時的寫法。在字母裡, 先 X 再 Y , 先有 i 才有 j , 先有 m 才有 n 。因此

$$\begin{array}{ll} X & \text{用足標 } i \text{ 計有 } m \text{ 個} \\ Y & \text{用足標 } j \text{ 計有 } n \text{ 個} \end{array}$$

其中隱含的次序, 數學上一點都不重要。但如此寫將出來, 會讓讀者在唸這一段時少用很多心力。並且你自己做計算時也會多費心力。例如 (4.2) 中的分母自然是 m , 不會是 n 。如果我們將 (4.1) 寫成

$$\begin{array}{ll} X_i = x_i, & X_i \sim_{\text{ind}} F, \quad i = 1, 2, \dots, m \\ Y_i = y_i, & X_i \sim_{\text{ind}} F, \quad i = 1, 2, \dots, n \end{array}$$

雖然數學上正確, 但就不夠漂亮了。即使我們寫成

$$\begin{array}{ll} X_i = x_i, & X_i \sim_{\text{ind}} F, \quad i = 1, 2, \dots, m \\ Y_k = y_k, & X_k \sim_{\text{ind}} F, \quad k = 1, 2, \dots, n \end{array}$$

也不夠漂亮。這類的小小化妝, 也是要學的。

p.12, section 5, line 7. 這一節目的是建立 bootstrap 和 jackknife 的關係, 是和論文標題有關的主要結論。在 line 7 處, 就讚揚 [10] 是 excellent paper。此事世間已有公論, 所以誇講一下也無所謂¹¹。何況 bootstrap 還要用 jackknife 來做墊背用! 墊背夠好, 被墊的文章當然會更好。

這一段的數學技巧是將 X^* 看成為多項分布。這是 standard trick, 因為原來的 $X_i \sim F$ 用到 X_i^* , 就變成了 P^* , 而後者的分布是可以寫出 exact form 的。技術上, 等於是由一個 non-parametric 的架構, 轉換到 parametric 的架構上去¹²。

另一個要點是 R 在這一段中完全沒有限制。因此此段說的是: 一般的 bootstrap 和一般的 jackknife 間的關係。當然, (5.11) 所得到的結果, 和 [10] 中所得一絲不差——這才是本文中較硬的結果。此事到了 p.14 的第一段結束, 才算真正回到傳統的 jackknife。

這一段要唸得真懂, 得找張紙找枝筆跟著做。此外, 需得去讀 [10]¹³ 或者 [14]。

這一節以 2006 年的角度來看, 並不重要——因為 bootstrap 已可以獨立站起。但本文的標題是 “Bootstrap methods: Another look at the jackknife”, 因此, 「借用 jackknife 以成就本文」的味道很濃。而這一節就變得有意義。任何情形下, jackknife 和 bootstrap 都是重抽法 (resampling method), 因此, 在第一篇介紹 bootstrap 的文章中, 「和 jackknife 比一比」, 「和 jackknife 拉上關係」都是有意思的。這一節雖然對後來使用 bootstrap 的人無用, 但對於 referee 的影響, 卻是有一點用的——至少, 這一節讓我們對於 bootstrap 和 jackknife 都有較深的了解, 而了解卻是研究的目的。

本節的最後一段有點多餘, 只是回應前面的對稱情形, 再多晃一招而已¹⁴。最後, (5.15) 中最後一行中 $\bar{y}$ 的足標是多餘的。

p.14, section 6. 這一節又舉了一個例子, 例子性質又與以前的例子不同: (1) 它做的是標準的 2-sample case, 因此可望由此看出一般的迴歸該怎麼做¹⁵。(2) 對於 2-sample

¹¹但對不怎麼樣的論文, 就別亂捧: 可能有人要吃味, 也可能讓評審人覺得你品味不足。

¹²嚴格地講, 因為還有 $x_1, x_2, \dots, x_n$ 的值可視為有 n 個未知參數的 parametric model。

¹³現在大概是找不到它了。因為 Bell Labs 已變成 Lucent, 而後者差不多要垮了。

¹⁴我不敢說一定是虛晃一招。

¹⁵結果迴歸未明顯地由此看出。

問題而言, 尚有不知如何處理 jackknife 的問題存在。而經由 bootstrap 的角度以觀之, 可能解決 jackknife 這個問題。(3) 我們可算出一些 exact 的結果 [公式 (6.7)], 亦可經由 method 2 (delta method) 算出近似的結果 [公式 (6.13)], 而這兩者相差極少 —— 從而說明 delta method 用在 bootstrap 上也是有道理的。

在 p.16, (6.14) 之後, 明白說明如何用 jackknife 來做一個 2-sample 的問題, 這解決了一個未決的問題。

p.17, section 7. 對這一節我的感覺不好。主要的是其技術內容, 和前面的 2-sample case 似無交集, 也不搭調。也許從 Wilcoxon test 直接到 regression 是太快了, 我看不出之間的關連。(2.7) 固然解釋了「做法正確」, (7.8) 又用 jackknife 來墊背 —— 說 [15] 和 [9] 都做錯了。但這一節因為和前面連不上, 寫來是有一點弱的。

當然, 文章到了尾巴這裡, 文氣已盡。讀者到此也都累了, 一般的想法是: 前面的似乎都沒有錯, 最後一節大概也錯不了, 因此,「深究」大概是不會去做的了。

p.19, section 8. Remark 指的是文章中較有深度的註脚。此文有 6pp 的 remarks, 是長了些。但這是 Efron 的風格。

總的來說, 這文章其實只有一個主要的 idea: 「若我們想知道 $R(\mathbf{X}, F)$ 的分布, 用 bootstrap 就好」。除了幾個有解析解的地方外¹⁶, 全文其實沒有多少技術面的東西。但這卻是極好的統計論文: 見林又見樹, 而見林的部分, 遠大於見樹, 因為作者選了一個廣大的角度來看 bootstrap 的問題。數學計算太多的統計論文, 基本上都是令人卻步, 只有一小撮人可能有興趣的。

這文章沒有講甚麼? 想一想。有沒有模糊的地方需要弄清楚的? 哪些問題他該問卻沒有問?...

這是「找題目來做」的方法之一。當然, 這是老文章了, 容易補充的地方, 大概都被補齊了。去好好查一下 1979 後出現的關於 bootstrap 的文章, 我想, 只是讀它們的緒論, 你就該學會怎樣找題目了。

September 4, 2006

¹⁶而且都不算難。

第八部分：評審幾篇論文

本部分是個習題。附上四篇論文，請你用 referee 的角度來讀它。你的態度是「儘量找麻煩」，因為你預備建議主編「這文章不能要」。因此你要挑出所有的錯來。

爲省空間，我們拿掉所有的圖。

第一篇（此文的摘要和緒論曾用作習題）

Testing for Activation in Data from FMRI Experiments

Abstract: The traditional method for processing functional magnetic resonance imaging (FMRI) data is based on a voxel-wise, general linear model. For experiments conducted using a block design, where periods of activation are interspersed with periods of rest, a haemodynamic response function (HRF) is convolved with the design function and, for each voxel, the convolution is regressed on prewhitened data. An initial analysis of the data often involves computing voxel-wise two-sample t-tests, which avoids a direct specification of the HRF. Assuming only the length of the haemodynamic delay is known, scans acquired in transition periods between activation and rest are omitted, and the two-sample t-test is used to compare mean levels during activation versus mean levels during rest. However, the validity of the two-sample t-test is based on the assumption that the data are Gaussian with equal variances. In this article, we consider the Wilcoxon rank test as well as modified versions of the classical t-test that correct for departures from these assumptions. The relative performance of the tests are assessed by applying them to simulated data and comparing their size and power; one of the modified tests (the CW test) is shown to be superior.

Key words: Excess kurtosis, haemodynamic response function, Shapiro-Wilk test, skewness, two-sample t-test, Welch test, Wilcoxon Rank test.

1. Introduction

Functional Magnetic Resonance Imaging (FMRI) is a non-invasive method that produces a time sequence of images of a subject's brain that are sensitive to changes in blood oxygenation caused by neural activation. The vast majority of analytical techniques that are applied to FMRI data assume the transfer function between neural activation and subsequent changes in blood oxygenation, the haemodynamic response function (HRF), is known fully *and* the data follow the Gaussian distribution. In this article, we consider the analysis of FMRI data collected in one of two states, called "activation" and "rest," based on two-sample tests. From knowledge of the length of the haemodynamic delay, measurements during the transition period between activation and rest can be omitted. The validity of the classical two-sample t-test is based on the assumption that the activation data and the rest data are Gaussian with equal variances. In this article, we propose

use of a modified two-sample test for FMRI data that allows for departures from this assumption. We study three competing tests. One is the Welch test (Welch, 1937), which is a modification of two-sample t-test that allows unequal covariances. A second competitor is the Cressie-Whitford (CW) test (Cressie and Whitford, 1986) that can be used with non-Gaussian data. The third competitor is the Wilcoxon rank (WR) test (Wilcoxon, 1945). In what follows, we compare the classical t-test with the Welch, CW, and WR tests for FMRI data based on a block design, where the blocks alternate between periods of activation and rest.

The next section describes the physiological background and physical processes used in FMRI and the most common methods used to process FMRI data; it also defines the four two-sample tests (including the classical two-sample t-test) that are compared in Section 4. Section 3 discusses the application of the two-sample tests for FMRI data and describes the methods used to identify and quantify departures from Gaussianity for each voxel. The size and power of the four tests are compared in Section 4 using a simulation study of FMRI data, from which recommendations are given. Section 5 contains discussion and conclusions.

2. FMRI Experiments

2.1 Some physiology

All neuronal activation is linked to an increase in oxygen consumption, causing a local increase in the blood flow. The body's response is to supply more oxygen than is required for the neuronal activity. Due to the different magnetic properties of oxygenated and de-oxygenated blood, the excess oxygenated blood that circulates during neuronal activation alters the magnetic properties of the venous blood, resulting in the so-called *blood oxygenation level dependent* (BOLD) signal. FMRI produces a sequence of brain images that is sensitive to changes in the BOLD signal.

In a classical FMRI experiment, the subject is scanned every few seconds to obtain an image of the brain; the subject is exposed to an experimental stimulus in some time periods, and is in a rest state during the remaining time periods. The stimulus can either be applied for brief periods in rapid, possibly random succession (an "event-related" experimental design, Josephs *et al.*, 1997), or for longer periods with interspersed rest periods (a "block" experimental design, Frackowiak *et al.*, 1997). In this paper, we focus on FMRI experiments conducted using a block experimental design.

Even though neuronal activation occurs immediately after exposure to the experimental stimulus, the vascular response evolves more slowly, resulting in the BOLD signal. The temporal relationship between neuronal activation and the observed BOLD signal is called the haemodynamic response. To model the haemodynamic response, it is common to convolve the experimental design with a so-called haemodynamic response function (HRF). Poisson, gamma, and Gaussian distributions are used widely as HRFs (Friston *et al.*, 1994).

The region of the brain where there is neural activation is found by regressing the

observed fMRI data on the expected BOLD signal, obtained as a convolution of the experimental design with the HRF. Of course, this depends on a well-specified HRF.

2.2 fMRI data

Observed fMRI data are four-dimensional, in space and time. At each time point, a three-dimensional image of the brain is acquired, called a volume. Each volume consists of voxels, and each voxel has an associated one-dimensional time series of observed signal intensities.

The most common approach to the analysis of fMRI data is to consider the voxels independently. A widely-used approach assumes a general linear model (GLM) for the voxel-wise time series (Friston *et al.*, 1995). For example, after various preprocessing steps, including prewhitening to achieve approximately independent errors, a two-sample test statistic is computed for each voxel where the two samples correspond to activation data and rest data. A voxel is declared to be significant if the test statistic exceeds some threshold. The distribution theory associated with this approach is based on the assumption of Gaussianity of the observed data and the proper specification of the HRF leading to the expected BOLD signal.

For initial data analysis, it is enough for us to know the length of the haemodynamic *delay* between neural activation and changes in the BOLD signal (Bandettini *et al.*, 1993). This knowledge is used to omit scans acquired in transition periods between possibly “activated” BOLD signals and “resting” BOLD signals. The delay between the neural activation and changes in the BOLD signal depends on many different factors; the type of stimuli, the duration of each stimulus, and the brain activation regions can all effect the length of the delay. Empirical studies have proposed methods for estimating HRFs that can adapt to different experimental designs. By using the block designs described in Section 2.1 and deleting transition data in our preliminary analysis, we have a sample of data acquired under activation and a second sample of data acquired under rest. In the next section, we describe four possible two-sample tests that might be used to test for the presence of activation at each voxel.

2.3 Two-sample tests

The null hypothesis of no difference between the means of two populations can be investigated with appropriate two-sample tests. In what follows, we summarize the four tests to be compared where, under activation the voxel data are $\mathbf{F}Y_a = \{Y_i\}_{i \in \mathbf{F}A}$ and, under rest the voxel data are $\mathbf{F}Y_r = \{Y_j\}_{j \in \mathbf{F}R}$; here $\mathbf{F}A$ and $\mathbf{F}R$ denote the activation and rest acquisition times, respectively.

The classical two-sample t-test

The classical two-sample t-test assumes:

(A1) Observations $\mathbf{FY}_a$ and $\mathbf{FY}_r$ are uncorrelated.

(A2) The observations within each of $\mathbf{FY}_a$ and $\mathbf{FY}_r$ have identical Gaussian distributions; that is,

$$\mathbf{FY}_f \sim \text{Gau}(\mu_f * \mathbf{F}1, \sigma_f^2 * \mathbf{F}I); \quad f \in \{a, r\}.$$

(A3) $\sigma_a^2 = \sigma_r^2$.

To test the hypothesis:

$$H_0: \mu_a \leq \mu_r \text{ versus } H_1: \mu_a > \mu_r, \quad (2.1)$$

the classical two-sample t-test uses test statistic,

$$T \equiv \frac{\bar{Y}_a - \bar{Y}_r}{\sqrt{\left(\frac{1}{n_a} + \frac{1}{n_r}\right) \left(\frac{(n_a-1)s_a^2 + (n_r-1)s_r^2}{n_a + n_r - 2}\right)}}, \quad (2.2)$$

with

$$\bar{Y}_f = \frac{1}{n_f} \sum_{i \in \mathbf{F}} Y_i \quad \text{and} \quad s_f^2 = \frac{\sum_{i \in \mathbf{F}} (Y_i - \bar{Y}_f)^2}{n_f - 1}; \quad f \in \{a, r\},$$

where $\mathbf{F}$ is the set of activation times $\mathbf{FA}$ (rest times $\mathbf{FR}$) if $f = a$ ($f = r$), and n_a (n_r) is the number of the observations in the sample $\mathbf{FY}_a$ ($\mathbf{FY}_r$).

If Assumptions (A1), (A2), and (A3) are satisfied, the classical two-sample t-test with significance level α is:

$$\begin{aligned} &\text{Accept } H_0 \text{ if } T < t_d(1 - \alpha) \\ &\text{Accept } H_1 \text{ otherwise,} \end{aligned}$$

where $t_d(1 - \alpha)$ is the $100(1 - \alpha)$ percentile of the t distribution on $d = n_a + n_r - 2$ degrees of freedom.

The Welch test

The Welch test (Welch, 1937) is used to test the same hypotheses (2.1), but it assumes only (A1) and (A2); that is, it is possible that $\sigma_a^2 \neq \sigma_r^2$. Welch (1937) has shown that under the null hypothesis H_0 , the test statistic

$$T^* \equiv \frac{\bar{Y}_a - \bar{Y}_r}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_r^2}{n_r}}} \quad (2.3)$$

has approximately a t distribution with

$$e \equiv \frac{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)}{\left(\frac{\sigma_a^4}{n_a^2(n_a-1)} + \frac{\sigma_r^4}{n_r^2(n_r-1)}\right)} \quad (2.4)$$

degrees of freedom. In practice, the population variances σ_a^2 , σ_r^2 in (2.4) are estimated from data using sample variances s_a^2 , s_r^2 . The Welch test with significance level α is:

$$\text{Accept } H_0 \text{ if } T^* < t_e(1 - \alpha)$$

Accept H_1 otherwise,

where the cut-off value $t_e(1 - \alpha)$ is based on fractional degrees of freedom and is obtained by interpolation of the $t_d(1 - \alpha)$ cut-off levels based on the nearest integers d to e .

The CW test

The CW test (Cressie and Whitford, 1986) also tests hypotheses (2.1), but makes only Assumption (A1); that is, it is possible that the data are non-Gaussian with unequal variances. To account for this, we use the same statistic T^* given by (2.3) as Welch, but modify its null distribution according to the skewnesses α_{3a} , α_{3r} and the excess kurtoses α_{4a} , α_{4r} of the non-Gaussian activation and rest distributions, respectively.

By calculating the Cornish-Fisher expansion of T^* , Cressie and Whitford (1986) show that under Assumption (A1) and H_0 , the distribution of T^* is approximately that of the random variable,

$$V = U + \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{6\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^{3/2}}(U^2 - 1) - \frac{\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}}{2\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^{3/2}}U^2 - \frac{1}{2}gUZ, \quad (2.5)$$

where U, Z are i.i.d. $N(0, 1)$ and

$$g \equiv \left\{ \frac{\frac{\sigma_a^4}{n_a^3}(\alpha_{4a} + 2) + \frac{\sigma_r^4}{n_r^3}(\alpha_{4r} + 2)}{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^2} - \frac{\left(\frac{\alpha_{3a}\sigma_a^3}{n_a^2} - \frac{\alpha_{3r}\sigma_r^3}{n_r^2}\right)^2}{\left(\frac{\sigma_a^2}{n_a} + \frac{\sigma_r^2}{n_r}\right)^3} \right\}^{1/2}. \quad (2.6)$$

The CW test with significance level α is

Accept H_0 if $T^* < v(1 - \alpha)$

Accept H_1 otherwise,

where $v(1 - \alpha)$ is the $100(1 - \alpha)$ percentile of the distribution of V , obtained by simulation. As for the Welch test, the population moments in (2.5) and (2.6) are estimated from data using sample versions; see Section 3.3.

The Wilcoxon Rank (WR) Test

The WR test (Wilcoxon, 1945) makes only assumption (A1), as does the CW test. In addition, it assumes that the distribution function $F(y)$ of the observations $\mathbf{FY}_r$ is continuous and the distribution function of the observations $\mathbf{FY}_a$ is $F(y - \delta)$, for $\delta \in \mathbb{R}$. Then the WR statistic tests the hypotheses,

$$H_0 : \delta \leq 0 \text{ versus } H_1 : \delta > 0. \quad (2.7)$$

In order to test (2.7), the WR test sums the ranks of each of the $\mathbf{FY}_a$ values in the combined sample of $N = n_a + n_r$ data consisting of the $\mathbf{FY}_a$ and $\mathbf{FY}_r$ values ordered from smallest to largest. Let R_i denote the rank of Y_i ; $i \in \mathbf{FA}$. The test statistic for the WR test is

$$W = \sum_{i \in \mathbf{FA}} R_i.$$

An exact p -value is then computed based on the null distribution ($\delta = 0$) of W , which is obtained by considering all possible $N!$ permutations of ranks of the $\mathbf{FY}_a$ and $\mathbf{FY}_r$.

However, this is computationally demanding for large n_a and n_r . For large n_a and n_r , we approximate the distribution of the centered and scaled version of W ,

$$W^* = \frac{W - .5 - n_a(n_a + n_r + 1)/2}{\sqrt{n_a n_r (n_a + n_r + 1)/12}},$$

with a standard normal (Hollander and Wolfe, 1999). Hence the WR test with significance level α is:

$$\begin{aligned} &\text{Accept } H_0 \text{ if } W^* < z(1 - \alpha) \\ &\text{Accept } H_1 \text{ otherwise,} \end{aligned}$$

where $z(1 - \alpha)$ is the $100(1 - \alpha)$ percentile of the Gaussian distribution with zero mean and unit standard deviation.

3. Methods of Analysis and Comparisons

In this section, we continue to consider inference based on a single generic voxel. Simultaneous inference involving all voxels is considered in Section 4.

3.1 Application of Two-Sample Tests to FMRI Data

Let $\mathbf{T}$ be the set of acquisition times of the observed intensities associated with the given voxel. Assuming the subject was exposed to only one type of neural activation, $\mathbf{T}$ can be divided into three groups: the time points $\mathbf{FA}$ where activation of the BOLD signal is expected, the time points $\mathbf{FR}$ during which the BOLD signal is expected to be in a rest state, and the time points $\mathbf{B}$ corresponding to the transition periods between the activation and the rest times. An example of such a division of time points is illustrated in Figure ???. In the two-sample tests considered in this article, one sample corresponds to $\mathbf{FA}$ and other sample corresponds to $\mathbf{FR}$; intensities corresponding to $\mathbf{B}$ are omitted from further analysis.

Figure 1 about here

Consider the two-sample tests of H_0 versus H_1 given in Section 2. For a given voxel and a given test, accepting the alternative hypothesis H_1 means that the associated voxel is declared to be activated by the experimental stimulus.

3.2 Simulated FMRI data

Six datasets were obtained from 3 healthy volunteers (1 female, 2 males) using a 1.5T Signa scanner. The data were collected under rest conditions; that is, the subjects were not exposed to any stimulus during the experiment and they were instructed to relax in the scanner with their eyes closed. One such rest dataset was obtained from the first male subject (30 years old), two rest datasets were obtained from the second male subject (27

years old), and three rest datasets were obtained from the female (30 years old). Each dataset consisted of 200 volumes, every observed volume contained 28 slices, and each slice had 64×64 voxels. These datasets were preprocessed for motion correction and prewhitened to make the time series uncorrelated (using the software FEAT, which is part of the FSL package; see Smith *et al.*, 2001).

We created activation datasets by essentially adding a signal having *known magnitude and location* of the activation to each preprocessed rest dataset. The signal component was calibrated against an image acquired from a previous unrelated visual-activation fMRI experiment; see Figure ?? for an example. By applying the signal in the locations acquired from a previous visual experiment, we avoided the possibility of applying the signal near so-called default regions (regions which show decreased neuronal activity during the activation of the stimulus) and their confounding effects on the simulated signal. The activation datasets alternated blocks of 10 time points of rest with 10 time points of activation. The average peak-signal change, defined as a ratio between the average of the intensities under the activation and the average of the intensities measured during the rest periods for the most activated voxel, was set to be 3%. Each dataset contains 200 time points; the three sets of time points $\mathbf{FA}$, $\mathbf{FR}$, and $\mathbf{B}$ were obtained assuming a haemodynamic delay of 3 time periods, resulting in $n_a = 70$ and $n_r = 73$.

Figure 2 about here

3.3 Violations of equal variances and Gaussianity assumptions

Several methods were used to assess the degree of departure of the activation datasets from (A2) and (A3). Consider a generic voxel and recall from Section 1 that $\mathbf{FY}_a = \{Y_i\}_{i \in \mathbf{FA}}$ make up the so-called “activated” sample and $\mathbf{FY}_r = \{Y_j\}_{j \in \mathbf{FR}}$ make up the “rest” sample.

To investigate the violation of Assumption (A3) given in Section 2, thereby allowing $\sigma_r^2 \neq \sigma_a^2$, we computed the sample variances for $\mathbf{FY}_a$ and $\mathbf{FY}_r$ for each voxel in each activation dataset. The pairs of sample variances of active and rest samples for all voxels that are located in subject’s brain (out of all $64 \times 64 \times 28 = 114,688$ voxels, only 2,2340 of them were located in subject’s brain) are plotted in Figure ??; the 45-degree line corresponding to equal variances is superimposed. In all panels, and especially in 3(c), we see some points far from the diagonal, which suggests that the assumption of homogeneity is violated for three voxels. A formal F-test ($\alpha = 0.05$) of equal variances detected 1,225 out of 22,340 (5.5%) brain voxels to have significantly different sample variances, and visual inspection of these voxels indicated no spatial pattern. This indicates that, overall, unequal variances may not be a serious problem for these fMRI data.

Figure 3 about here

To investigate departures from Gaussianity, Assumption (A2), we computed the sample skewness and sample excess kurtosis for $\mathbf{FY}_a$ and $\mathbf{FY}_r$, for all six activation datasets. For

the activation sample these are:

$$\begin{aligned}\hat{\alpha}_{3a} &= \frac{\sqrt{n_a} \sum_{i \in \text{FA}} (Y_i - \bar{Y}_a)^3}{\{\sum_{i \in \text{FA}} (Y_i - \bar{Y}_a)^2\}^{3/2}}, \\ \hat{\alpha}_{4a} &= \frac{n_a \sum_{i \in \text{FA}} (Y_i - \bar{Y}_a)^4}{\{\sum_{i \in \text{FA}} (Y_i - \bar{Y}_a)^2\}^2} - 3,\end{aligned}$$

and likewise we computed $\hat{\alpha}_{3r}$ and $\hat{\alpha}_{4r}$ for the rest sample.

To illustrate graphically the relationship between skewness and kurtosis, we chose one activation dataset. The pairs $(\hat{\alpha}_{3a}, \hat{\alpha}_{4a})$ for the 22,340 brain voxels from one activation dataset are plotted on the left panel of Figure ??, and the pairs $(\hat{\alpha}_{3r}, \hat{\alpha}_{4r})$ are plotted on the right panel. For Gaussian data, the plotted pairs should be very close to the origin. In Figure ??, we observe strong departures from zero skewness and zero excess kurtosis in both panels. Thus, we might expect an improvement in hypotheses testing for activation using the CW test or the WR test over the classical two-sample t-test or the Welch test.

Figure 4 about here

More formally, we calculated the Shapiro-Wilk test (e.g., Royston, 1982) for normality ($\alpha = .05$) for each voxel and rest/activation combination. For the dataset used in Figure ??, Table ?? summarizes the number (out of 22,430) of brain voxels that were significantly non-Gaussian. About 12% of activated samples and about 11% of rest samples were declared significant by the Shapiro-Wilk test; if the samples were Gaussian, we would expect only 5% to be declared significant. More than 20% of voxels were declared significant in at least one of the activated or rest samples.

Table 1: Brain-voxels declared significant using Shapiro-Wilk test ($\alpha = .05$), based on one of the six datasets.

		Activated samples		Total
		Significant	Not significant	
Rest samples	Significant	647	2095	2742 (12.3%)
	Not significant	1774	17824	19598
Total		2421 (10.8%)	19919	22340

The spatial distribution of the voxels declared significant is shown in Figure ??; while they are distributed fairly homogeneously between regions of the brain, there is some indication that, within a region, they can clump together.

Figure 5 about here

4. Results

All four two-sample tests were used to test for activation in each voxel. We obtained p-values as in Section 2 where the p-value for the CW test was obtained from simulation of the random variable given by (2.5) and that for the WR test was obtained from the standard normal approximation to W^* .

Because of the multiple hypotheses being tested (one for each brain voxel), the voxels declared as active were obtained by comparing the p-values with $\alpha^* \equiv \alpha / \{\# \text{ of brain voxels}\}$ with $\alpha = .05$. This is the voxelwise Bonferroni-adjusted level of significance based on an overall level of significance of $\alpha = .05$. Voxels with p-values less than or equal to α^* were pronounced active. Because the activation pattern of each dataset was known, we can estimate and compare the sizes and powers of the two-sample tests.

Let A denote the set of voxels to which an activation signal has been added and R the set of voxels with no added activation. Let $\mathcal{A}_{\text{right}}$ denote the voxels in A declared to be active, and let $\mathcal{A}_{\text{wrong}}$ denote the voxels in A not declared active. All voxels from category R can be similarly divided into $\mathcal{R}_{\text{right}}$, those non-activated voxels not declared active, and $\mathcal{R}_{\text{wrong}}$, those non-activated voxels which were declared active.

The achieved size of each test was estimated by

$$\hat{\alpha} \equiv (|\mathcal{R}_{\text{wrong}}| / |\mathcal{R}|),$$

where $|\mathcal{C}| \equiv \#$ voxels in the region $\mathcal{C}$ of the brain. The quantity $\hat{\alpha}$ is also called the false-positive rate and should be comparable to the desired familywise level of significance α ($= .05$). If $\hat{\alpha} < \alpha$, the test is conservative. The power of each test was estimated by

$$\hat{\pi} \equiv (|\mathcal{A}_{\text{right}}| / |\mathcal{A}|),$$

which is the true-positive rate.

Table ?? lists the estimated sizes and powers of each test for all six simulated FMRI datasets. All four tests were consistently very conservative, with the Wilcoxon test being the most conservative. The classical t-test and Welch test had equivalent power, which was consistently greater than that of the Wilcoxon test. The CW test was the most powerful test, uniformly over the six datasets.

Table 2: Estimated size and power of the four two-sample tests for the six datasets.

Dataset	TEST							
	Classical t-test		Welch		CW		WR	
	$\hat{\alpha}$	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\pi}$	$\hat{\alpha}$	$\hat{\pi}$
1	.496E-4	.289	.496E-4	.288	.992E-4	.305	0	.277
2	0	.208	7.466E-4	.208	19.927E-4	.224	4.479E-4	.200
3	0	.221	0	.217	0	.235	0	.202
4	.583E-4	.233	.583E-4	.233	1.750E-4	.253	.583E-4	.224
5	0	.205	0	.205	0	.223	0	.188
6	0	.239	0	.237	27.527E-4	.251	1.101E-4	.227

Table ?? gives a more detailed comparison of the classical t-test and the CW test for one of the datasets. While 626 out of 2,173 activated brain voxels were correctly detected as significant by both tests, 37 additional activation voxels were correctly detected by the CW test that were not identified by the classical t-test. Only one activation voxel was identified by the classical t-test that was missed by the CW test.

Table 3: Comparison of the performance of the CW test and the classical t-test, based on one of the six datasets

			CW test			
			Voxels from $\mathcal{A}$		Voxels from $\mathcal{R}$	
			$\mathcal{A}_{\text{right}}$	$\mathcal{A}_{\text{wrong}}$	$\mathcal{R}_{\text{right}}$	$\mathcal{R}_{\text{wrong}}$
Classical t-test	Voxels from $\mathcal{A}$	$\mathcal{A}_{\text{right}}$	626	1	.	.
		$\mathcal{A}_{\text{wrong}}$	37	1509	.	.
	Voxels from $\mathcal{R}$	$\mathcal{R}_{\text{right}}$	.	.	20165	1
		$\mathcal{R}_{\text{wrong}}$	.	.	0	1

5. Discussion and Conclusions

While the results were obtained from only one type of scanner, the 1.5T Signa GE, and with fMRI data for three subjects, they show that fMRI data can exhibit both unequal variances and non-Gaussianity. Using the Shapiro-Wilk test, more than 20% of voxels in the dataset were declared significant in one or both of the rest or activated samples. We believe that more powerful scanners will lead to data that are even more non-Gaussian, since their finer spatial resolution involves less averaging of the response.

The Welch test is valid for unequal variances but when non-Gaussianity is suspected, the CW test accounts for both. The WR test is a nonparametric analog of the classical t-test. In the six datasets studied in Section 3, non-Gaussianity was a bigger problem than unequal variances. The results in Section 4 showed that the CW test performed better than the other three tests. These results suggest that the CW test should replace any standard use of the classical parametric or nonparametric two-sample tests based on fMRI data.

Acknowledgement

This research was supported by the Office of Naval Research under grants N00014-99-1-0214 and N00014-02-1-0052 and by the National Science Foundation Grant DMS-0406026. The authors would like to thank Antonio Algaze and Petra Schmalbrock for providing the fMRI data and the members of FMRI, Oxford UK for initial consultation about simulating activation fMRI datasets. Perceptive comments by the referees led to improvements in the exposition and strengthening of our conclusions.

References

- Bandettini, P. A., Jesmanowicz A., Wong, E. C. and Hyde, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine* **30**, 161-173.
- Cressie, N. and Whitford, H. J. (1986). How to use the two sample t-test. *Biometrical Journal* **28**, 131-148.
- Frackowiak, R. S. J., Friston, K. J., Frith, C. D., Dolan, R. J. and Mazziotta, J. C. (1997). *Human Brain Function*. Academic Press.
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C. D. and Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* **2**, 189-210.
- Friston, K. J., Jezzard, P. and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping* **2**, 69-78.
- Hollander and Wolfe, (**** add initials ****) (1999). *Nonparametric Statistical Methods*, 2nd edn. John Wiley and Sons.
- Josephs, O., Turner, R. and Friston, K. J. (1997). Event-related fMRI. *Human Brain Mapping* **5**, 243-248.
- Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* **31**, 115-124.
- Smith, S. M., Bannister, P., Beckmann, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, J., Lebovici, D., Ripley, B., Woolrich, M. and Zhang, Y. (2001). FSL: New tools for functional and structural brain image analysis. *NeuroImage* **13**, S249.
- Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

第二篇 (此文的摘要和緒論曾用作習題)**An Evaluation of Multiple Behavioral Risk Factors for Cancer in a Working Class, Multi-Ethnic Population**

Abstract: Behavioral risk factors for cancer tend to cluster within individuals, which can compound risk beyond that associated with the individual risk factors alone. There has been increasing attention paid to the prevalence of multiple risk factors (MRF) for cancer, and to the importance of designing interventions that help individuals reduce their risks across multiple behaviors simultaneously. The purpose of this paper is to develop methodology to identify an optimal linear combination of multiple risk factors (score function) which would facilitate evaluation of cancer interventions.

Key words: Community based research, conditional logistic regression, multiple risk factors, random effects.

1. Introduction

Despite the considerable biomedical advances of the last half-century, facilitating improvement in lifestyle behaviors remains the most efficacious population-level strategy for reducing cancer risk. Estimates vary, but suggest that over fifty percent of new cancer cases and up to one-third of cancer mortality could be prevented through improvements in health behavior practices (American Cancer Society, 2004; Doll and Peto, 1981). A 19 percent decline in the rate at which new cancer cases occur, and a 29 percent decline in the rate of cancer deaths, could potentially be achieved by 2015, if prevention efforts were heightened and behavior change sustained. This would translate to the prevention of approximately 100,000 cancer cases and 60,000 cancer deaths each year, by the year 2015 (National Cancer Policy Board and Institute of Medicine, 2003).

There is ample epidemiological evidence for the consideration of red meat consumption, physical activity, and folic acid intake in cancer prevention efforts. Regular physical activity lowers the risk of cancers of the colon, breast, and possibly prostate (Colditz, Cannuscio, and Frazier, 1997; Friedenreich and Rohan, (1995).). An additional 30 percent of cancer deaths can be attributed to adult diet (Anonymous, 1996); higher intake of red meat has been associated with increased risk of colon (Sandhu, White and McPherson, 2001) and prostate cancers (Michaud, Augustsson, Rimm, Stampfer, Willett, and Giovannucci 2001). Associated with both physical inactivity and diet is obesity, which may account for between 25-30 percent of cancers of the colon, breast (postmenopausal), endometrium, kidney, and esophagus (Vainio and Bianchini, 2002). Folic acid is protective against colon cancer (Giovannucci, Stampfer, Colditz, Hunter, Fuchs, Rosner, Speizer, and Willett, 1998); long-term multi-vitamin use, in particular has been found to reduce risk for colon cancer, likely because of its folic acid content (Giovannucci *et al.* 1998).

The risk for many diseases, including colon cancer, is associated with multiple behavioral risk factors (MRF); these behaviors are highly interrelated and tend to cluster within individuals. For example, those who eat high-fat diets are also more likely to be sedentary, suggesting that the behaviors may be mutually reinforcing (see e.g., Emmons, Marcus, Linnan, Rossi, and Abrams, 1999). Change in one behavioral risk factor thus may serve as a stimulus or gateway for change in the other health behaviors (see e.g., Emmons *et al.* 1999), and there are overarching behavioral principles and intervention frameworks that guide behavior change efforts across risk factors.

Consequently, to facilitate population-level reductions in cancer risk, it may be inefficient to target discrete behavioral risk factors, when similar principles might be applied simultaneously to multiple behaviors (Institute of Medicine, 2000).

The literature provides little consensus as to the most appropriate analytic strategy for evaluating the efficacy of MRF interventions; most studies have analyzed the various outcomes independently or by creating a simplistic sum (e.g., 1 RF + 1 RF = 2RFs) (see e.g., Prochaska and Sallis 2004; Campbell, James, Hudson, Carr, Jackson, Oakes, Demissie, Farrell, and Tessaro, 2004). This could be problematic, because the use of separate analytic strategies may result in improper inferences regarding the effect of an MRF intervention because of correlation among the factors. Such strategies may overlook the clustering effect brought about by the agglomeration of multiple behavioral risk factors and have been criticized as being too simplistic. The purpose of this paper is to develop a methodology to identify an optimal linear combination of multiple behavioral risk factors (MRF score function) for cancer that would best facilitate evaluation of an MRF cancer intervention.

2. Methods

2.1 Study design

The data analyzed in this paper are from the Harvard Cancer Prevention Program Project (HCPPP) Healthy Directions, which is composed of two randomized controlled trials, one in health centers (HC) (Emmons, Stoddard, Gutheil, Suarez, Lobb, and Fletcher 2003), and another in small businesses (SB) (Hunt, Stoddard, Barbeau, Wallace, and Sorensen 2003). The overarching goal of the HCPPP was to create a new generation of cancer prevention interventions that would be effective among working class, multi-ethnic populations. Together, the two arms of the trial were successful in enrolling a sub-population of the multi-ethnic working class population in eastern Massachusetts. The study aims and sampling strategies are published in greater detail elsewhere (Emmons *et al.*, 2003; Hunt *et al.*, 2003).

2.2 Health centers

Healthy Directions-HC (Emmons *et al.*) was a randomized controlled trial conducted in collaboration with a large health care delivery system, comprised of 14 multi-specialty medical group practices that serve over 270,000 patients. Ten of the fourteen health centers were invited to participate in this study, and all agreed. Health center served as the unit of randomization and intervention. Briefly, patients who resided in low income, multi-ethnic neighborhoods (defined using census block-groups that were predominantly working class, impoverished, or with low levels of education) were identified and approached for participation through their health center. Individuals identified through geocoding to be residents in the target neighborhoods were deemed eligible if they met the following criteria: (1) being 18-75 years old, (2) having a well-care or follow-up visit scheduled with a participating provider, (3) being able to speak and read either English or Spanish, (4) not having cancer at the time of enrollment, (5) not being employed by the participating health centers, (6) not being employed by a worksite participating in the companion small business study, and (7) providing consent to participate in the randomized study. All providers practicing in the Internal Medicine Departments of the health centers were approached for permission to recruit from among their patient pools. Provider participation averaged 83% across sites (range 50%-100%; 97 clinicians). Patients scheduled for appointments with the participating providers and in the eligible age range were identified through the automated central appointment system. Study staff attempted to recruit 8,963 potentially eligible candidates; 2,547 (28%) individuals were

unreachable. Among the 6,414 potential subjects reached, 867 (14%) were ineligible, 3,330 (52%) refused, and 2,219 (35%) were enrolled. Assuming that 14% of those not reached were also ineligible, the response rate is 29% of those assumed eligible. The cohort recruited at baseline was contacted by telephone after the intervention period to complete a follow-up survey. Of the 2,219 who completed the baseline survey (n=1088 intervention condition; n=1131 control condition), 1,954 (88%) completed the follow-up survey. The follow-up response rate was equivalent across conditions.

2.3 Small business

The Healthy Directions-SB study (Hunt *et al.*, 2003) was a randomized controlled trial in which the worksite was the unit of randomization and intervention. Worksites were identified using the Dun and Bradstreet database to locate small businesses with Standard Industrial Classification (SIC) codes 20-39 (manufacturing industries) and employing between 50-150 employees. Additional inclusion criteria included: (1) employing a multi-ethnic population (defined as 25% of workers being first-or second-generation immigrants or people of color), (2) having a turnover rate of less than 20% in the previous year, (3) being autonomous in decision-making power to participate in a study, and (4) agreeing to be randomly assigned to the intervention condition. One hundred thirty-three (133) companies met the eligibility criteria, and of these, 26 agreed to participate (Barbeau, Wallace, Lederman, Lightman, Stoddard and Sorensen 2004).

Data were collected using interviewer-administered surveys among individuals who were permanent employees and worked 20 hours or more per week. On site interviews were administered on company time in the language (either English, Spanish, Portuguese, or Vietnamese) preferred by respondents. Two cross-sectional samples were collected, one at baseline in which 1,740 participants from 26 worksites completed the survey (response rate 84%). The second sample was collected at follow-up 1,408 participants in 24 worksites (during the course of the intervention two worksites dropped out, one intervention and one control) with a response rate of 77%. 974 participants (518 in control worksites and 456 in intervention worksites) completed both the baseline and follow-up surveys forming the embedded cohort used in this analysis.

2.4 Data and analysis

The goals of the intervention were to: (1) increase fruit and vegetable intake, (2) decrease red meat consumption, (3) increase physical activity levels, and (4) increase daily multivitamin usage. The following variables assess the individual risk factors measured on a continuous scale: number of servings of fruit and vegetables per day, number of servings of red meat consumed per week (RM), and hours of moderate or vigorous physical activity per week (PA). The fourth measure is a binary variable indicating use of a multi-vitamin on 6 or 7 days per week (MV). In order to keep all variables on an equivalent time scale, we created a new variable for fruit and vegetable consumption that calculated the amount of fruits and vegetables consumed in one week (FV) by multiplying the current measure of fruit and vegetable intake by seven. The continuous variables (FV, RM, PA) were standardized using the formula in Equation 2.1;

$$STV = \frac{V - P_{05}}{P_{95} - P_{05}}, \quad (2.1)$$

where V are the original values for the continuous variables (FV, RM, PA), P_{05} and P_{95} are the fifth and ninety-fifth percentile values respectively for a given variable and STV are the new standardized variables (STFV, STRM, and STPA respectively). Standardization was implemented for consistency (to make a one unit change in one variable similar to a one unit change in another) and interpretability. The 5th and 95th percentiles were used to minimize the influence of outliers.

For the purposes of identifying an optimal linear combination that would show an intervention effect we restricted our sample to only those subjects who received the intervention, responded to both the baseline and follow-up surveys, and have complete data for the four risk behaviors. As opposed to the usual situation of observing how the covariate vector or a linear combination of the covariate vector will change because of treatment, the idea here is to determine how the covariate vector or the linear combination will predict the intervention status. This is similar in spirit to a matched case-control analysis.

A popular method for the analysis of longitudinal data with a dichotomous outcome is a mixed effects logistic regression model. A mixed effects logistic regression model with a logit link will have the form:

$$\log \left[\frac{\text{pr}(Y_{ij} = 1)}{1 - \text{pr}(Y_{ij} = 1)} \right] = a_i + \beta' X_{ij}. \quad (2.2)$$

Here Y_{ij} , $i = 1, \dots, n$, $j = 1, 2$, denotes the indicator of intervention time (i.e. pre-intervention $Y_{i1} = 0$ and post-intervention $Y_{i2} = 1$), X_{ij} is the covariate vector, and a_i is a random cluster effect. The subscript i is an indicator for individual and the subscript j is an indicator for time. Each individual subject i is a cluster of two sets of observations, pre-intervention and post-intervention. The random effect variable a_i can be thought of as measuring an individual's demographic characteristics (i.e., age, gender, race). In our analysis, we want to control for an individual's specific demographic characteristics, therefore, we treat the random effect variable a_i as a nuisance parameter and condition it out of the model. We can condition them out by using the conditional likelihood based on the fact that $Y_{i1} + Y_{i2} = 1$. We are left with a conditional logistic regression model. These types of models are often used to analyze matched case-control studies, where the outcome of interest is whether a subject is a case or control.

In this framework we intend to model

$$\text{logit}(\text{Pr}(Y_{ij} = 1|a_i)) = \beta' X_{ij}, \quad (2.3)$$

where an optimal linear combination, or the best score, will be $\hat{\beta}'X$.

We set up our data as if it came from a 1:1 matched case-control study; each individual is a cluster of two observations, one "case" and one "control". One observation is pre-intervention ("control"/baseline) and the second observation is post-intervention ("case"/follow-up). At each time point (pre and post-intervention) each subject has a vector (containing STFV, STRM, STPA, and MV) of covariates.

For matched case-control studies with one case per matched set, the likelihood function for the conditional logistic regression reduces to the partial likelihood of the Cox model for the continuous time scale (Hosmer and Lemeshow 1998). We created dummy survival times so that all cases have the same event time and the corresponding controls are censored at a later time. We used Proc PHREG in SAS¹ to fit the conditional logistic regression model by forming a stratum for each matched set (individual id number). This allowed us to obtain estimates for $\hat{\beta}$.

3. Results

Using the combined Health Center and Small Business data from the Healthy Directions baseline and follow-up surveys on the 1,209 study participants that received the intervention, we found an optimal score function for the four risk factors:

$$\text{score} = 1.05 * \text{STFV} + 1.70 * \text{MV} + 0.25 * \text{STPA} - 1.35 * \text{STRM}. \quad (3.1)$$

The score is a summary measure of the health behaviors of a subject based on these four factors. From this score, we can see that increasing the number of fruits and vegetables consumed per week,

taking a multivitamin six or more days a week, increasing the amount of physical activity done in a week, and/or decreasing the amount of red meat consumed in a week will increase the score for a subject which in turn means an overall improvement in health behaviors. The dynamics of the score are consistent with the goals of the intervention. A participant can increase their health behavior score by changing one risk factor, or combinations of the four risk factors in a manner consistent with the goals of the intervention.

We believe that these factors not only have individual effects, but that some factors may also have compounding effects. This belief is based on previous evidence of the interrelationships seen in modifying behavioral risk factors (see e.g., Emmons *et al.*, 2004; Butterfield *et al.*, 2004). Therefore, we looked for significant interactions between the four variables. Table ?? shows the analysis of maximum likelihood estimates for our final model. In our final score function (see Equation 3.2), we multiply the effects (parameter estimates) by 100 to increase the range of the scores as well as to simplify interpretation.

Table 1: Analysis of Maximum Likelihood Estimates

Standardized Variable	Parameter Estimate	Standard Error	P-Value
STFV	0.576	0.303	0.0570
MV	2.008	0.2078	<.0001
STPA	0.232	0.193	0.2294
STRM	-1.515	0.343	<.0001
STFV*STRM	1.229	0.565	0.0296
MV*STRM	-0.707	0.343	0.0392

$$\begin{aligned} \text{score} = & 57.6 * STFV + 200.8 * MV + 23.2 * STPA - 151.5 * STRM \\ & + 122.9 * STFV * STRM - 70.7 * MV * STRM \end{aligned} \quad (3.2)$$

There was a significant interaction between the amount of fruits and vegetables consumed per week and the amount of red meat consumed per week, suggesting that changing both behaviors simultaneously is better than changing either behavior alone, but the effect of changing both behaviors is not equal to the sum of the individual changes on the MRF score. There was also a significant interaction between multivitamin usage more than six times a week and the amount of red meat consumed per week, suggesting that changing either behavior alone is good, but changing both behaviors simultaneously will result in an even larger increase on the MRF score.

Table 2: Examples of changes in individual risk factor measures and resulting MRF score

	FV	MV	PA	RM	MRF Score
Baseline values	20	0	4	5	-30.23
<i>Case 1: Optimal values at final</i>					
Final values	35	1	10	1	247.46
change	+15	+1	+6	-4	+277.69
<i>Case 2: Improves only FV</i>					
Final values	35	0	4	5	17.39
change	+15	0	0	0	+47.62
<i>Case 3: Improves only MV</i>					
Final values	20	1	4	5	135.22
change	0	+1	0	0	+165.45
<i>Case 4: Improves only PA</i>					
Final values	20	0	10	5	-19.09
change	0	0	+6	0	+11.14
<i>Case 5: Improves Only RM</i>					
Final values	20	0	4	1	14.64
change	0	0	0	-4	+44.87
<i>Case 6: Improves FV and RM</i>					
Final values	35	0	4	1	72.82
change	+15	0	0	-4	+103.05

Table ?? displays a few examples of how a change in an individual risk factor from the baseline case to the optimal case will change the score. If we consider the first row of Table ?? to be a baseline value in which a subject consumes 20 servings of fruits and vegetables per week, does not take a multivitamin six or more days a week, has four hours of physical activity per week, and consumes five servings of red meat per week (the average values for study subjects at baseline, meeting only the recommend level of physical activity), the standardized values would be 0.32, 0.32, and 0.5 respectively. Therefore the score for a subject at baseline would be

$$score = 57.6 \cdot 0.32 + 200.8 \cdot 0 + 23.3 \cdot 0.32 - 151.5 \cdot 0.5 + 122.9 \cdot 0.32 \cdot 0.5 - 70.7 \cdot 0 \cdot 0.5 = -30.2. \quad (3.3)$$

We can consider an arbitrary optimal case as a subject who consumes 35 servings of fruits and vegetables per week (or five a day), takes a multivitamin 6 or more days a week, engages in 10 hours of physical activity per week, and eats one serving of red meat per week (meeting and/or exceeding all of the recommended levels). Table ?? shows the effects of these changes on the score from the baseline case to the optimal case for each variable alone and the effects of combinations of two and three variables. Figure ?? compares our final model (MRF, Equation 3.2) with a main effects model (a model without interactions) showing that the main effects model can both overestimate and underestimate scores predicted from the MRF model due to the absence of the two significant interactions.

Table 3: Score changes with one, two, and three variable changes

Variables Changed	Score Change
FV	47.62
MV	165.45
PA	11.14
RM	44.87
FV + MV	213.07
FV + PA	58.76
FV + RM	72.82
MV + PA	176.59
MV + RM	238.60
PA + RM	56.00
FV + MV + PA	224.21
FV + MV + RM	266.55
FV + PA + RM	83.96
MV + PA + RM	249.73

Although we used only those subjects that received the intervention to develop the score, the score is generalizable to the entire study population. It was created, and is most useful for, the purpose of comparing the subjects that received the intervention to those that received usual care, because it provides a summary measure of the health behaviors of a subject on all intervention risk factors pre and post-intervention. There were 1,297 subjects that received usual care and took both the baseline and follow-up surveys. These subjects can be considered controls for the effect of the intervention. Figure ?? shows box plots of score comparing baseline and follow-up for subjects that received the intervention compared to those that received usual care. In the intervention group, the mean score at baseline was 48.1, while the mean score at follow up was 104.3. In the usual care group the mean score at baseline was 40.4, and the mean score at follow-up was 53.2. The mean change in score for the usual care group was 12.8, while the mean change in score for the intervention group was 56.2. There was a statistically significant difference in the mean change in score from baseline to follow-up when comparing the usual care group to the intervention group ($p < 0.001$). The intervention group showed greater improvements in score at follow-up proving the intervention quite successful.

4. Discussion

Increasing attention has been paid to multiple risk factor interventions, across a range of disease outcomes, both because adverse behavioral risk factors tend to cluster within individuals and because of recognition of the utility of facilitating change across multiple risk behaviors. However, most MRF studies to date have used individual risk factor methods to analyze intervention effects (see e.g., Prochaska and Sallis (2004); Campbell *et al.*, 2004). As shown in Figure ??, the main effects model both over-estimates (e.g., FV & PA & RM) and under-estimates (e.g., MV & PA & RM) the scores predicted from the MRF model, depending on the combination of variables and the degree of change for a given participant in the intervention. Thus, such analytic models may compromise determinations of the efficacy of a MRF intervention. We were successful in modeling a linear combination of behavioral risk factors including interactions between risk factors, an effort that represents an advance over the existing methods for analyzing MRF intervention efficacy.

To illustrate, note that in our final model there are two interaction terms. One between the amount of fruits and vegetables consumed per week and the amount of red meat consumed per week, and another between multivitamin usage more than six times a week and the amount of red

meat consumed per week. Looking at Table ??, we can see that with all the other variables held constant, a change in fruit and vegetable consumption alone from 20 to 35 servings per week will increase the score by 47.62, and a decrease in red meat alone from 5 to 1 servings per week will increase the score by 44.87. However, because of the interaction term, if both variables are changed by the amounts indicated the score would increase by 72.82, which because of the interaction is a smaller than 92.49, the sum of the individual changes. Similarly, if a subject begins to take a multivitamin daily the score will increase by 165.45, and if they decrease red meat from 5 to 1 serving per week the score will increase by 44.87. However, if a participant begins to take a multivitamin daily and decreases red meat consumption by 4 servings per week the score will increase by 238.60, a larger increase than 210.32 that you would get by adding 165.45 from taking a multivitamin daily and 44.87 by decreasing red meat consumption. Cluster effects are not captured by main effects models and are an advantage of this method.

There are some limitations to the method proposed here, namely that the score function depends on the efficacy of the intervention to determine variable weighting. For example, if the intervention was most effective at increasing multivitamin use, the weight (coefficient) for the multivitamin use variable would be largest in magnitude, whereas if the intervention was least effective in changing the participants' physical activity patterns, the weight (coefficient) for the physical activity variable would be the smallest in magnitude. In some cases then, the weights may be a proxy for the amount of participant effort necessary to change the health behavior. For example, in this study we saw that multivitamin usage had the largest weight and thus the most influence on the score.

There are at least two potential explanations for this finding. First, the promotion of multivitamin usage may require less participant burden when compared to the other health behaviors (e.g., physical activity). Thus, it may be easier for participants to modify their multivitamin use; this supposition appears to be supported by the finding of an almost thirty percent increase from baseline to follow-up of the number of subjects taking a multivitamin daily. However, it is important not to undermine the significance of a change in multivitamin usage which is strongly related to the prevention of disease outcomes. Sustained use of multivitamins containing folic acid have been associated with the reduction in risks for numerous conditions including colorectal cancer and cardiovascular disease (Giovannucci *et al.*, (2002); Fairfield and Fletcher, 2002). Physical activity on the other hand, is among the most challenging health behaviors addressed in the study to intervene upon. In this population, 66 percent of the subjects were getting the recommended level of physical activity at baseline, and 69 percent at follow-up. Of those subjects that were not at or above the target level of physical activity at baseline, almost 9.5 percent were at or above the target level at follow-up. Another factor to consider is that multivitamin usage was treated as a binary variable in our models. That is, many potential changes are captured in the categorization of either taking a multivitamin 6 or more times a week or not doing so. Relative to increasing one serving of fruits and vegetables a week, decreasing one serving of red meat in a week, or increasing an hour of physical activity a week, this is a substantial change.

Although the purpose of our method was to develop a health behavior score (composite variable), there are some limitations to using this type of variable. The purpose of such a variable is to allow for easy comparisons of the four factors with one number. When there are changes in the score, however, a composite variable does not provide any insight into which individual risk factor(s) have contributed to the change.

Another potential limitation of applying this method to the HCPPP data is the merging of the two cohorts, small businesses and health centers. Our method develops a score function that is independent of the population but not independent of the intervention. By combining the two data sets, we have made the assumption that the interventions given to these two populations are the same. In reality, although the two interventions were quite similar, they were not identical. We decided, however, to combine the two cohorts in order to increase power, and to create a universal score that could be applied to both cohorts. This not only allowed us to make comparisons within

a cohort, but between cohorts. Taking these limitations into account, our methodology remains preferable compared to existing techniques that do not accord weights to the risk factors or adjust for cluster effects.

In summary, we have developed a score that effectively integrates multiple behavioral cancer risk factors into one measure, irrespective of individual demographic factors. We believe that the methods are generalizable to other working class multi-ethnic populations, and future research should be done to evaluate the effectiveness of these methods in other groups. The primary strength of the methodology used to develop the score is that it can be easily implemented to develop scores for other populations, for other combinations of behavioral risk factors, or for other disease outcomes (e.g., cardiovascular disease). Given the increasing attention being paid to the development of MRF interventions, we believe the described method to be the preferred means of analysis in comparison to previously used strategies. Ultimately, we believe that analytic focus on examining clusters of behavioral risk factors will enhance the design of multiple risk factor intervention approaches.

Figure 1 about here

Figure 2 about here

Acknowledgements

The research of Melody Goodman was supported by National Institute of Child Health and Human Development grant 5 F31 HD043695. The research of Dr. Li was supported by National Institute of Health grant R01CA95747. The research of Dr. Bennett, Dr. Stoddard, and Dr. Emmons was supported by grant 5 P01 CA75308 from the National Institutes of Health, and support to the Dana-Farber Cancer Institute by Liberty Mutual, National Grid, and the Patterson Fellowship Fund.

The statistical output for this paper was generated using SAS/STAT software, Version 8 of the SAS System for Windows. Copyright© 1999-2001 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

References

- American Cancer Society (2004). Cancer Facts and Figures. Technical Report, American Cancer Society.
- Anonymous (1996). Harvard Report on Cancer prevention Volume: 1 Causes of Human Cancer, *Cancer Causes and Control* **7**, S3-S9.
- Barbeau, E. M., Wallace, L., Lederman, R., Lightman, N., Stoddard, A. and Sorensen, G. (2004). Recruiting small manufacturing worksites that employ multi-ethnic, low-wage workers to a cancer prevention research trial. *Preventing Chronic Disease* **1** 1-9.
- Butterfield, R. M., Park, E. R., Puleo, E., Mertens, A., Gritz, E. R., Li, F. P., and Emmons, K. (2004). Multiple risk behaviors among smokers in the childhood cancer survivors cohort. *Psychooncology* **13**, 619-629.

- Campbell, M. K., James, A., Hudson, M. A., Carr, C., Jackson, E., Oakes, V., Demissie, S., Farrell, D. and Tessaro, I. (2004). Improving multiple behaviors for colorectal cancer prevention among african american church members, *Health Psychol* **23**, 492-502.
- Colditz, G. A., Cannuscio, C., and Frazier, A. (1997). Physical activity and reduced risk of colon cancer: Implications for prevention. *Cancer Causes and Control* **8**, 649-667.
- Doll, R. and R. Peto (1981). The causes of cancer : quantitative estimates of avoidable risks of cancer in the United States today, *J. Natl. Cancer Inst.* **66**, 1191-1308.
- Emmons, K. M., Marcus, B. H., Linnan, L. A., Rossi, J. S. and Abrams, D. B. (1999) Mechanisms in multiple risk factor interventions: Smoking, physical activity, and dietary fat intake among manufacturing workers. *Preventive Medicine* **23**, 481-489.
- Emmons, K. M., Stoddard, A. M., Gutheil, C., Suarez, E. C., Lobb, R. and Fletcher, R. (2003). Cancer prevention for working class, multi-ethnic populations through health centers: The healthy directions study. *Cancer Causes and Control* **14**, 727-737.
- Fairfield, K. M and Fletcher, R. H. (2002). Vitamins for chronic disease prevention in adults: scientific review. *Journal of the American Medical Association* **287**, 3116-3126.
- Friedenreich, C. M., and Rohan, T. E. (1995). Physical activity and risk of breast cancer. *European Journal of Cancer Prevention* **4**, 145-151.
- Giovannucci, E., Stampfer, M. J., Colditz, G., Hunter, D., Fuchs, C., Rosner, B., Speizer, F. and Willett, W. (1998). Multivitamin use, folate, and colon cancer in women in the Nurse's Health Study. *Annals of Internal Medicine* **129**, 517-524.
- Hosmer, D. W., and Lemeshow, S. (1998). *Encyclopeida of Biostatistics*, 2327-2333. John Wiley.
- Hunt, M. K., Stoddard, A., Barbeau, E. M., Wallace, L. and Sorensen, G. (2003). Cancer prevention for working class, multiethnic populations through small businesses: The Healthy Directions Study. *Cancer Causes & Control* **14**, 749-760.
- Institute of Medicine (2000). *Promoting Health: Intervention Strategies from Social and Behavioral Research*. National Academy Press.
- Michaud, D. S., Augustsson, K., Rimm, E. B., Stampfer, M. J., Willett, W. C. and Giovannucci, E. (2001). A prospective study on intake of animal products and risk of prostate cancer. *Cancer Causes & Control* **12**, 557-567.
- National Cancer Policy Board and Institute of Medicine (2003). Fulfilling the potential of cancer prevention and early detection, Technical Report, Washington D. C.
- Prochaska, J. J. and Sallis, J.F. (2004). A randomized controlled trial of single versus multiple health behavior change: promoting physical activity and nutrition among adolescents. *Health Psychology* **23**, 314-318.
- Sandhu, M. S. and White, I. R. and McPherson, K. (2001). Systematic review of the prospective cohort studies on meat consumption and colorectal cancer risk: A meta-analytical approach. *Cancer Epidemiology, Biomarkers & Prevention* **10**, 439-446.
- Vainio, H. and Bianchini, F. (2002). *IARC Handbooks of Cancer Prevention. Volume 6: Weight Control and Physical Activity*. IARC Press.

第三篇 (此文的摘要和緒論曾用作習題)

Reducing Subjectivity in the Likelihood

Abstract: Some scientists prefer to exercise substantial judgment in formulating a likelihood function for their data. Others prefer to try to get the data to tell them which likelihood is most appropriate. We suggest here that one way to reduce the judgment component of the likelihood function is to adopt a mixture of potential likelihoods and let the data determine the weights on each likelihood. We distinguish several different types of subjectivity in the likelihood function and show with examples how these subjective elements may be given more equitable treatment.

Key words: Mixture likelihood, model averaging, subjectivity.

1. Introduction

We propose methods for modeling the likelihood function that will require fewer subjective judgments. We first discuss the nature of the problem of subjectivity in the likelihood function; then we review some related research; and finally, we define a mixture likelihood function and suggest estimation procedures that reduce the effects of subjective views imposed on the observed data.

1.1 Statement of the problem

It is sometimes desirable that beliefs of experimenters should be brought into a scientific analysis in ways that minimally distort the measured data (see, for example, Hogarth, 1980; Kyberg and Smokler, 1980; Lad, 1996). But that having been said, scientists observing data sometimes interpret the data points subjectively, according to what they want the data to show, and according to how precisely they believe the data points were measured. The latter procedure is of course quite common. This subjective interpretation of observed data may be totally at the unconscious level, or it may be purposeful (with the purposeful interpretation, the analysis may become fraudulent; see for example, Grayson, 1995, 1997; Howson and Urbach, 1990; and Press and Tanur, 2001).

The subjective interpretation of empirical data in medicine was discussed by Kaptchuk (2003). He stated (page 1, *op. cit.*):

Doctors are being encouraged to improve their critical appraisal skills to make better use of medical research. But when using these skills, it is important to remember that interpretation of data is inevitably subjective and can itself result in bias. Facts do not accumulate on the blank slates of researchers' minds, and data simply do not speak for themselves. Good science inevitably embodies a tension between the empiricism of concrete data and the rationalism of deeply held convictions. Unbiased interpretation of data is as important as performing rigorous experiments. This evaluative process is never totally objective or completely independent of scientists' convictions or theoretical apparatus.

Statistical analysis of a data set most often proceeds by summarizing the distribution of the data in terms of its likelihood function. In order to specify the form of the likelihood function, various assumptions are made about the data, such as mutual independence, identical distributions, unimodality, etc. After the likelihood function has been specified, additional assumptions are sometimes made (significance levels thought to be appropriate are specified, a prior distribution about the underlying unobservable quantities may be brought in, etc.). Analysis of the data generally proceeds by trying to keep the likelihood function treatment of the data as simple as possible, so that the scientist or analyst will introduce minimal distortion of the data. The analyst tries not to discard data, and tries to maximize the chance of understanding what nature is trying to tell us through the revealed data about the underlying phenomenon. In this way, when the analysis of the data has been completed, the claim can reasonably be made that the conclusions drawn from the analysis approximate, if not precisely reflect, the laws of nature, rather than the possible misinterpretations and misunderstandings of the laws of nature by human beings. It will be useful to first briefly define what we mean by objectivity and subjectivity, in this context.

According to Mandik (2001)¹,

The word *objectivity* refers to the view that the truth of a thing is independent from the observing subject. The notion of objectivity entails that certain things exist independently from the mind, or that they are at least in an external sphere. Objective truths are independent of human wishes and beliefs. The notion of objectivity is especially relevant to the status of our various ideas, and the question is to what extent objectivity is possible for thought, and to what extent it is necessary.

This is but one of many definitions that have been suggested. The elusive quest for objectivity in science has been, and remains, an important topic of discussion among historians and philosophers of science (for extensive additional discussions of the meaning of “objectivity”, see for example, Bower, 1998; Porter, 1995, 1996; and Daston and Galison 1992). For some, scientific objectivity involves the search for *certainty* in knowledge about one of nature’s well-kept secrets, independent of what human beings believe; but in many cases, we find that what we earlier thought to be true about nature, turns out later to be questionable.

In an interesting example from physics, Folger, 2003, pointed out that:

Pioneer 10, launched in 1972, is now some 8 billion miles from home. But is has been slowing down, as if the gravitational pull on it from the sun is growing progressively stronger the farther away it gets. Milgrom proposed (see the MOND pages-MODified Newtonian Dynamics)² that Newton’s laws might change at these accelerations. If Milgrom is right, Newton’s and Einstein’s laws will be in for some major tweaking.

Sometimes the scientist has such deep understanding and insight into the phenomenon he/she is studying that the scientist’s own predictions of what should be found from the analysis are far superior to what the data analysis seems to indicate. In some cases the beliefs of the scientist or analyst are so strong, even before actually taking any data that bear on the phenomenon, that the data are interpreted or manipulated so that they will reflect these preconceived views of the scientist. Any preconceived personal views (views held before taking any data), weak or strong, are what we refer to in this context as *subjectivity*.

¹Mandik, P. (2001). *The Internet Encyclopedia of Philosophy*, <http://www.utm.edu/research/iep/o/objectiv.htm>.

²MOND pages — <http://www.astro.umd.edu/ssm/mond/>

1.2 Related Research

One approach to reducing the effects of differing assumptions about likelihoods may be found in a line of research that involves use of the *empirical likelihood function*. In this approach, most useful in large samples, a discretized, binned, version of the empirical cdf, instead of a specific likelihood function, is used. Inference is then made from a multinomial distribution. An unfortunate feature of this approach is the additional unknown parameters that are concomitantly introduced into the model. See: Owen, 1988, 2001. For typically small and moderate size samples this could be a problem, but for the massive data sets typical of data mining applications (see, for example: Berry and Linoff (1997); and Hastie, Tibshirani, and Friedman, 2001) such an approach could be a helpful alternative.

We show in the next section how we might understand and account for some types of subjectivity that sometimes enters the *likelihood function*, and might not be desired. We will use the definition and form of the likelihood function in which for absolutely continuous random variables, up to a proportionality constant, it is the joint probability density function of the observables given the unobservables.

2. Types of Subjectivity in the Likelihood Function

We distinguish three of the types of likelihood subjectivity problems that may occur:

- (a) how to determine the distributional form of the likelihood function in a way that is largely objective, but permits the data themselves to guide the modeling as to whether the data are Normally-distributed, or Gamma-distributed, or possibly follow some other convenient distribution. We call this problem, “*distributional subjectivity*”;
- (b) how to treat observed data that have possibly been weighted subjectively so that some data points are valued more heavily than others, and some are even ignored; we call this problem, “*weighted-data subjectivity*”;
- (c) how to account for the nature of the experiment used to obtain the data that may have favored one type of response over another; we call this problem, “*experiment subjectivity*”.

We treat each of these types of subjectivity in Section 3.

3. Reducing Likelihood Subjectivity

3.1 The mixture likelihood

We use a convex mixture of various likelihoods for the data; the usual likelihood function results as a special case.

Suppose an experiment is repeated n times with the resulting one-dimensional data outcomes: $x_1, x_2, \dots, x_n$. We suppose that there are J models for the data that potentially we might reasonably entertain. For simplicity, merely to suggest a general type of approach, we consider problems

involving only one unknown parameter, namely, the means of the J distributions, ω .

In some situations, the parameters may be quite different from one another but they can generally be related functionally. For example, the case of distinguishing between the means of normal and log-normal distributions, where the mean parameter has different meanings in the two cases is sometimes particularly interesting. In such cases, functional relationships among the parameters are required.

Suppose, in the one-parameter problem, we can assume these data to be mutually independent and identically distributed, and we agree to adopt the likelihood function for Model m_j :

$$\ell_j(x_1, \dots, x_n | m_j, \omega) \equiv \ell_j(\underline{x} | m_j, \omega).$$

Define a “mixture likelihood function”,

$$L_M(x_1, \dots, x_n | \omega) \equiv L_M(\underline{x} | \omega),$$

such that:

$$L_M(\underline{x} | \omega) = E\{\text{likelihood}\} = E_{Model}[\ell(\underline{x}) | \omega] = \sum_{j=1}^J \ell_j(\underline{x} | \omega) P(m_j | \omega) \quad (3.1)$$

where $\ell_j(\underline{x} | m_j, \omega)$ denotes the usual likelihood function of the data under model m_j , $\ell(\underline{x} | \omega)$ denotes a model-independent likelihood function, and $P(m_j | \omega)$ denotes the prior probability of model m_j . The mixture likelihood function is of course a likelihood function itself. If there were only one model ($J = 1$), L_M reduces to the ordinary likelihood. The mixture likelihood function explicitly assumes that we should combine different models in a linear way. Other possibilities exist of course, and perhaps in certain cases, they are even more desirable. But because for a wide variety of cases, the linear assumption seems appropriate, we will retain this assumption throughout. We next address the issue of how to reduce model subjectivity (how to choose the weights).

3.2 Reducing “model subjectivity”

In some instances, the scientist has very strong, theory-based, beliefs about how the data were generated, and how the corresponding likelihood function should behave. In such instances, especially in small samples, the analyst should surely use that information to permit the desired likelihood function to emerge. In other situations where the scientist/analyst wants the data to speak as loudly as possible relative to the scientist’s pre-conceived beliefs, there is no unique way to accomplish this objective. The approach suggested here is to take equal weights in the mixture. Accordingly, take all $P(m_j | \omega)$ in eqn. (3.1) to be equal (discrete uniform distribution). This interpretation of equal treatment for the different models is:

- (1) in keeping with the approach frequently used for weighting in mixture models to express indifference or ignorance among the various components in the mixture;
- (2) it is the procedure suggested by Laplace when he adopted his Principle of Insufficient Reason (Laplace, 1812, 1814);
- (3) it is consistent with a basic result of information theory that the distribution that corresponds to maximum entropy, or minimum information, is the uniform distribution.

This gives the mixture likelihood function (equally-weighted average likelihood):

$$L_M(\underline{x}|\omega) = \frac{1}{J} \sum_{j=1}^J \ell_j(\underline{x}|m_j, \omega). \quad (3.2)$$

For example, suppose there are just two potential models ($J = 2$) that might reasonably represent the data: $N(\omega, 1)$ and a Student t -distribution centered at ω , with 3 degrees of freedom (a fat-tail distribution that has a population mean). Then, the mixture likelihood function becomes:

$$L_M(\underline{x}|\omega) = \frac{1}{2} \left\{ \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \omega)^2\right\} + \prod_{i=1}^n \frac{m^{m/2}/B(1/2, m/2)}{[m + (x_i - \omega)^2]^{(m+1)/2}} \right\}. \quad (3.3)$$

Clearly each term in equation (3.3) is non-negative and integrates to one (with respect to $\underline{x}$), so $L_M\{\underline{x}|\omega\}$ is a bone fide likelihood function for the data (as would be the case whichever models we choose). In some situations, one scientist might favor the normal distribution for representing the distribution of the data, while another might favor the Student t -distribution. By using $L_M\{\underline{x}|\omega\}$ to represent the likelihood function for all inferences, the analyst reduces the model subjectivity in the description of the data distribution. Maximum likelihood estimation of ω is now more complicated numerically than it would be with use of either the normal or the Student t distributions separately, but the numerical problem is straightforward (see numerical example below) and easily generalizes to more than two possible ordinary likelihoods.

We next numerically illustrate the example suggested in this section of how to reduce model subjectivity when the models under consideration are the $N(\omega, 1)$ and the Student t_3 centered at ω . We randomly generated a total of 20 observations, 10 observations from t_3 , a Student t -distribution with 3 degrees of freedom centered at $x = 10$, and 10 observations from $N(10, 1)$. The resulting data are shown in columns 2 and 3 of Table 1a. Then, using the Newton-Raphson method, we calculated the mixed MLE. It is given at the bottom of Column 2 as: $\hat{\omega} = 9.9168$. To illustrate variability, there are four replications of this entire process shown in Table 1a; the four resulting mixed maximum likelihood estimates (mixed MLE's) are also shown in Table 1a.

Table 1a: Four replications of model subjectivity

	t	normal	t	normal	t	normal	t	normal
1	11.1861	9.6734	8.1266	11.182	8.9931	8.331	11.476	9.8325
2	9.9749	11.542	11.092	10.175	10.641	10.131	10.922	11.051
3	8.5632	10.259	11.374	11.720	9.8717	7.8108	10.270	10.642
4	5.5471	9.4442	9.3422	10.757	8.8824	8.3177	8.3906	9.0293
5	9.5188	10.779	13.189	9.8871	9.7579	9.4354	11.899	8.8359
6	10.994	9.3448	9.6007	9.715	10.385	10.092	9.5234	10.566
7	9.1875	9.9779	9.9069	9.7106	13.659	9.4326	14.559	9.495
8	11.283	9.2274	8.3785	9.8394	9.5543	10.361	10.177	10.247
9	8.7574	10.724	10.073	11.637	9.2285	9.0399	10.155	9.0938
10	7.9804	11.263	10.379	9.4837	11.274	9.7291	9.2795	10.366
Mixed MLE	9.9168		10.240		9.6237		10.116	

For comparison purposes, we also computed the separate ordinary MLE's assuming all 20 observations were generated from a normal, and then, that all 20 observations were generated from a t_3 distribution. Results are given in Table 1b.

Table 1b: Separate MLE's For normal and Student data

Normal MLE	9.761	10.278	9.7463	10.291
t_3 -MLE	9.924	10.187	9.6104	10.111

Thus, it may be seen that in the first instance, while the mixed MLE is 9.9168, the MLE assuming all 20 observations came from a normal is 9.7614, whereas the MLE assuming all 20 observations came from a t_3 is 9.9240. Results for the other 3 cases are shown in Tables 1a and 1b as well. Depending upon the assumptions made for the modeling, results for the mixture MLE obtained from the model averaging may differ substantially from those of the separate models, or not.

3.3 Reducing “weighted-data subjectivity”

We examine two distinct cases of weighted data subjectivity and model the two cases separately below.

Case 1 — Several Observers (Scientists) Rate the Same Data Points Differently

In this case, different observers (scientists) might interpret the same points differently. Some observers might view certain points as mistakes (outliers that were generated from different distributions from the other points), and therefore delete them from the analysis; and others might, according to their own beliefs, weight certain points more heavily than others (perhaps difficult-to-measure points might be weighted less heavily because the error associated with the measurement might be greater than with most of the other points; perhaps certain points obtained were measured under censored conditions; etc.).

For simplicity, assume the data points are mutually independent. We define the likelihood function for Observer O_k as:

$$\ell(\underline{x}|\omega) = \prod_{j=1}^n [f(\delta_{jk}x_j | O_k, \omega)]^{p_k(\delta_{jk}x_j|O_k)}, \quad (3.4)$$

where: $p_k(\delta_{jk}x_j|O_k) = 1$, if Observer O_k includes the data point x_j in the analysis, and $p_k(\delta_{jk}x_j|O_k) = 0$ if not; δ_{jk} denotes the weight that Observer k places on observation x_j , $f(x_j|\omega)$ denotes the pdf (probability density function) of X_j , conditional on ω . The mixture likelihood function may be defined as:

$$\begin{aligned} L_M(\underline{x}|\omega) &= E\{\text{likelihood}\} = E_{data}\{\ell_k(\underline{x}|\omega)\} \\ &= \sum_{k=1}^K \ell_k(\underline{x}|\omega) P_k(O_k), \end{aligned} \quad (3.5)$$

where $P_k(O_k)$ denotes the prior probability that the data analyst places on the model that has been developed by Observer k . To be objective (or indifferent among the choices), in the sense we have been discussing, we take $P_k(O_k) = 1/K$, for all k . Then,

$$L_M(\underline{x}|\omega) = \frac{1}{K} \ell_k(\underline{x}|\omega) \quad (3.6)$$

As a simple example, suppose that all K observers adopt the same distribution for the data, say, $N(\omega, 1)$ (in Section 3.2 the analyst adopted two different possible distributions for the data),

and assume that they weight the points in the same way, so that $\delta_{jk} = 1$ for all k , for all points they include in their analyses, but they may include different points. Then, since the n observations are independent,

$$\ell_k(\underline{x}|\omega) = \prod_{j=1}^n \left[\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_j - \omega)^2\right\} \right]^{p_k(x_j|O_k)}. \quad (3.7)$$

To be specific, suppose that $n = 102$, and that there are two observers, O_1 and O_2 . Suppose further that O_1 believes x_{102} is an outlier, and O_2 believes that both x_{101} and x_{102} are outliers, but they agree that the first 100 points $(x_1, \dots, x_{100})$ should be included in their analyses. Then,

$$\begin{aligned} p_1(x_j|O_1) &= 1 \quad \text{for } j = 1, 2, \dots, 101, \\ &= 0, \quad \text{for } j = 102, \end{aligned}$$

also,

$$\begin{aligned} p_1(x_j|O_2) &= 1 \quad \text{for } j = 1, 2, \dots, 100, \\ &= 0, \quad \text{for } j = 101, 102, \end{aligned}$$

Then,

$$\ell_1(\underline{x}|\omega) \equiv \ell_1 = \left(\frac{1}{\sqrt{2\pi}} \right)^{101} \exp\left\{-\frac{1}{2} \sum_{j=1}^{101} (x_j - \omega)^2\right\}, \quad (3.8)$$

and

$$\ell_2(\underline{x}|\omega) \equiv \ell_2 = \left(\frac{1}{\sqrt{2\pi}} \right)^{100} \exp\left\{-\frac{1}{2} \sum_{j=1}^{100} (x_j - \omega)^2\right\}, \quad (3.9)$$

Then,

$$\begin{aligned} L_M(\underline{x}|\omega) &= \frac{1}{2} \left\{ \left(\frac{1}{\sqrt{2\pi}} \right)^{101} \exp\left\{-\frac{1}{2} \sum_{j=1}^{101} (x_j - \omega)^2\right\} \right. \\ &\quad \left. + \left(\frac{1}{\sqrt{2\pi}} \right)^{100} \exp\left\{-\frac{1}{2} \sum_{j=1}^{100} (x_j - \omega)^2\right\} \right\} \end{aligned} \quad (3.10)$$

We may now estimate ω by maximizing $L_M(\underline{x}|\omega)$ with respect to ω . Note first that if we let n_1, n_2 be the numbers of data points used in the respective analyses of Observers O_1 and O_2 , they are also the numbers of terms in the two summations, and in this example, $n_1 = 101$ and $n_2 = 100$. We may readily find by ordinary differentiation, the mixture maximum likelihood estimator (mixture MLE) to be:

$$\omega = \alpha(\omega)\bar{x}_1 + [1 - \alpha(\omega)]\bar{x}_2, \quad 0 \leq \alpha(\omega) \leq 1, \quad (3.11)$$

where:

$$\alpha(\omega) \equiv \frac{n_1 \ell_1(\omega)}{n_1 \ell_1(\omega) + n_2 \ell_2(\omega)} \quad (3.12)$$

$$\bar{x}_1 \equiv \frac{1}{101} \sum_{j=1}^{101} x_j, \quad \bar{x}_2 \equiv \frac{1}{100} \sum_{j=1}^{100} x_j. \quad (3.13)$$

That is, we find the interesting result that $(\hat{\omega}|\ell_1, \ell_2)$ is a weighted average (actually a convex combination) of the separate MLE's that the two observers might adopt separately, and the weights

are their respective proportions of their ordinary likelihoods, an intuitively sensible result. But note that because $\alpha(\omega)$ depends upon ω , equations (3.11) and (3.12) must be jointly solved numerically for $\hat{\omega}$.

While in large samples, the (continuous) data will generally ultimately swamp any prior distribution weights placed on the data points (see Le Cam, 1956), in small or moderate size samples, certain very influential points that may have been deleted from an analysis can have substantial effects on the interpretation of the experiment outcomes.

We next illustrate this example numerically. We randomly generated 18 points from $N(0, 1)$. We then ordered the points, and added 2 larger outliers. We assumed the first observer dropped the largest point as an outlier, and the second observer dropped the two largest points as outliers. We then calculated the mixture MLE numerically from equation (3.11) using the Newton-Raphson method. We replicated the procedure four times to examine variation. Data are shown in Table 2a.

Table 2a: Four replications of weighted-data subjectivity

Observation	$N(10, 1)$	$N(10, 1)$	$N(10, 1)$	$N(10, 1)$
1	8.3959	7.6748	8.1260	7.7977
2	8.4063	7.8796	8.5249	8.8122
3	8.559	7.9954	9.6225	8.9922
4	8.7975	8.7684	9.6490	9.0079
5	9.3082	8.9002	9.7041	9.0501
6	9.6001	8.9819	9.7444	9.1783
7	9.8433	9.2957	9.7660	9.2580
8	9.9802	9.3553	10.0400	9.3645
9	10.2570	9.3687	10.1180	9.4404
10	10.5710	9.5069	10.3150	9.7344
11	10.6690	9.6790	10.4280	9.8685
12	10.6900	9.8179	10.5690	10.0880
13	10.7120	9.8868	10.5780	10.2120
14	10.7140	10.0860	10.6230	10.2380
15	10.8160	10.3790	10.6770	10.3900
16	10.8580	10.4620	10.7310	10.4440
17	11.1910	10.5510	10.7990	10.5690
18	11.2540	10.9440	10.8960	10.7810
19	12.0000	12.0000	12.0000	12.0000
20	13.0000	13.0000	13.0000	13.0000

Calculations of MLE's for the data in Table 2a are given in Table 2b:

Table 2b: MLE's for data with outliers

Mixture MLE	10.0406	9.4205	10.0568	9.6267
$\bar{x}_{18}$	10.0346	9.4185	10.0506	9.6237
$\bar{x}_{19}$	10.1380	9.5544	10.1532	9.7487

We see that for the data in column 2 of Table 2b, for example, the mixture MLE was 10.0406. Had the observers carried out separate MLE's, with Observer 1 dropping only the last observation, he would have found his MLE to be 10.1380, while Observer 2 who dropped both of the last 2 observations would have found her MLE to be 10.0346. While the differences are not large they are intended to be illustrative.

Case 2 — One Observer (Scientist) Rates Each Data Point Differently

The second case of weighted-data subjectivity involves a single scientist weighting the importance of the data points differently from one another. Here we envision a single scientist who has carried out an experiment many times, but sometimes, for one reason or another, the scientist carried out the experiment with extremely small error, whereas on some other occasions, the scientist associated the experimental outcomes with considerably more error. Thus, which observed results had small associated error, and which had large associated error might differ from one replication of the experiment to the next.

In this context there is just one scientist who rates his/her experimental data differentially, according to how "well" the data point was measured, or what he/she thought should have occurred, or whatever. This is the more typical situation, compared with the first case. The mixture likelihood function is obtained from equations (3.4) and (3.5), for $K = 1$, as:

$$L_M(\underline{x} | \omega) = \ell_1(\underline{x} | \omega) P_1\{O_1\} = \prod_{j=1}^n [f(\delta_{j1} x_j | O_1, \omega)]^{p_1(\delta_{j1} | O_1)} . \quad (3.14)$$

To follow the paradigm suggested here we should take $\delta_{j1} = 1$ for every j . Of course the individual scientist would often argue that he/she knows better than anyone else that certain points were really not as good as others, and should therefore be down-weighted.

A now-classical example of this type of subjectivity of special historical interest has been documented with real data. It involves the data collected by R. A. Millikan (1868-1953). Dr. Millikan was an American physicist who successfully measured the charge on a single electron, winning a Nobel Prize in 1923 for this famous oil-drop experiment (as well as other prizes). Holton (1978) scrutinized Millikan's laboratory notebooks and found that Millikan had repeated his oil-drop experiment 39 times, obtaining outcomes: $x_1, \dots, x_{39}$ for the charge on the electron. Holton reported that Millikan had given each of his original sets of observations a personal quality-of-measurement rating: "best", "very good", "good", "fair", and no rating at all for discarded measurements (we interpret his weights to represent his prior probabilities for these measurements). The distribution of his rating results is summarized in the Table 3.

Table 3: Millikan's measurements

rating descriptions	effective raating	δ_{j1} = Weight	number of measurements
best	4	4/10	2
very good	3	3/10	7
good	2	2/10	10
fair	1	1/10	13
discard	no rating	—	7

For Millikan, $p_1(\cdot) = 1$, for 32 data points and $p_1(\cdot) = 0$ for the discarded 7 points. We order the measurements according to their effective ratings, from “best” to “fair”, and form the weighted average. The estimated value of the charge on the electron is then given by the weighted average:

$$\hat{e} = \frac{4}{10} \sum_{j=1}^2 x_j + \frac{3}{10} \sum_{j=3}^9 x_j + \frac{2}{10} \sum_{j=10}^{19} x_j + \frac{1}{10} \sum_{j=20}^{32} x_j.$$

Millikan formed the weighted average of his measurements and accordingly estimated the charge on the electron as 4.85×10^{-10} esu (electrostatic units). The ordinary equally weighted average would have been 4.70×10^{-10} esu. In his reported value he also averaged in the values obtained by other researchers. By contrast, the accepted value for “ e ”, the charge on the electron, today, is 4.77×10^{-10} esu. But the impressive closeness of Millikan's values with today's accepted value is deceptive; it occurred only because his values were based upon, “a faulty value for the viscosity of air, which when corrected, increases the discrepancy with the modern value by over 40%” (Mathews, 1998).

3.4 Reducing “experiment subjectivity”

Suppose there are two experiments that might be performed: E_g (“ g ” for “good”), and $E_{\bar{g}}$ (“ $\bar{g}$ ” for “not good”). In E_g the scientist knows that the experiment will contain one or more variables that might produce effects that will be confounded with the effect of fundamental interest. In $E_{\bar{g}}$, there are likely to be fewer such confounding variables, so the scientist believes that he/she is more likely to be able to distinguish the effect he/she is seeking. Concomitantly, it may be that by carrying out E_g , the scientist is missing the important variables that suggest that the effect sought is really artifactual, and the seeming effect is explainable in other ways. Because the scientist is so convinced that the effect sought is real and not artifactual, he/she reasons that E_g is a “cleaner” and more promising experiment. The scientist might even argue, in a moment of enthusiastic zeal, that E_g is cheaper and/or less subject to error.

In both experiments, for simplicity of interpretation, we assume the data are normally distributed with variance equal to 1. Suppose that the scientist referred to above, call him/her Scientist A, would like to show that the population mean for the underlying phenomenon of interest is positive. If Scientist A carries out E_g , it is more likely that the sample mean $\bar{x}$ will be positive than if Scientist A carries out $E_{\bar{g}}$ wherein the sample mean $\bar{y}$ will imply the alternative hypothesis $H_{\bar{g}}$: that the population mean is not positive. If $E_{\bar{g}}$ is performed the scientist believes results are either unlikely to be supportive of the theory, or they are likely to be sufficiently marginal so that the theory will be in doubt. A priori, the experimenter adjudges the chances for concluding H_g : the population mean is positive, when performing E_g as greater than the chances for concluding that the population mean is positive when performing $E_{\bar{g}}$. Consequently, Scientist A decides to perform E_g .

Suppose some other scientist, say Scientist B, performs $E_{\bar{g}}$, and subsequently observes $\bar{y}$ (using the same sample size, n). Let θ denote an indexing parameter such that $\theta = 1$ if the hypothesis H_g is true, and $\theta = 0$ if the hypothesis $H_{\bar{g}}$ is false.

$$\begin{aligned} L_M\{\text{data} \mid \theta\} &= E\{\text{likelihood}\} = E_{\text{experiment}}[\ell(\text{data} \mid \theta)] \\ &= \ell(\bar{x} \mid E_g, \theta)P\{E_g\} + \ell(\bar{y} \mid E_{\bar{g}}, \theta)P\{E_{\bar{g}}\} \end{aligned}$$

The mixture likelihood function becomes:

$$L_M\{\text{data} \mid \theta\} = P\{E_g\} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\} + P\{E_{\bar{g}}\} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{y} - \theta)^2\right\}.$$

An investigator cognizant of both experiments has both available. In the same spirit of a desire for equity of treatment in the likelihood function, the investigator takes $P\{E_g\} = P\{E_{\bar{g}}\} = 0.5$. Then,

$$L_M\{\bar{x}, \bar{y} \mid \theta\} = \frac{1}{2} \left[\frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2\right\} + \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left\{-\frac{n}{2}(\bar{y} - \theta)^2\right\} \right].$$

Define $z = (\bar{x} + \bar{y})/2$. Then, combining terms shows that:

$$L_M\{z \mid \theta\} = \frac{1}{2} \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\{-n/4\} \exp\{-n(z - \theta)^2\}.$$

Thus, the MLE for θ is clearly: $\hat{\theta} = z = (\bar{x} + \bar{y})/2$. If Scientist A were correct in his/her a priori assessments of what was likely to happen in the experiment, $\hat{\theta}$ is likely to be closer to zero than $\bar{x}$ (or even negative), a result that would tend to vitiate Scientist A's conclusions.

For example, for Scientist A's experiment, E_g , we generated 100 observations from $N(1, 1)$ and found $\bar{x} = 1.0598$. Then, for Scientist B's experiment, $E_{\bar{g}}$, we generated 100 observations from $N(-1, 1)$ and found $\bar{y} = -.9531$. So the generalized MLE, $\hat{\theta}$, is 0.053, a sample value just barely positive, which might not be convincing in many contexts for asserting that the population mean is really positive.

4. Conclusions

We have been concerned with how to reduce the effects of a scientist's pre-conceived beliefs in the analysis of his/her supposedly objectively-observed data. We have found that we can reduce the effect of some of those subjective interpretations by using a mixture likelihood function, and then choosing the mixture weights that weigh the various interpretations of the data equally.

References

- Berry, M. J., and Linoff, G. (1997). *Data Mining Techniques*. John Wiley.
- Bower, B. (1998). Objective visions: Historians track the rise and times of scientific objectivity. *Science News* **154**, 360-362.
- Daston, L. J. and Galison, P. (1992). The image of objectivity. *Representations* **40**, 81-128.

- Folger, T. (2003). Nailing down gravity: New ideas about the most mysterious power in the universe. *Discover Magazine*, Oct., 2003, 34-41.
- Grayson, L. (1995). *Scientific Deception*. The British Library.
- Grayson, L. (1997). *Scientific Deception – An Update.*: The British Library.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.
- Hogarth, R. (1980). *Judgment and Choice*. John Wiley.
- Holton, G. (1978). Sub-electrons, presuppositions, and the Millikan-Ehrenhaft dispute. In *The Scientific Imagination: Case Studies* (Edited by Gerald Holton), 25-83. The Cambridge University Press.
- Howson, C. and Urbach, P. (1990). *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing Co.
- Kaptchuk, T. J. (2003). Effect of interpretive bias on research evidence. *British Medical Journal* **326**, 1453-1455.
- Kyberg, H. E. Jr., and Smokler, H. E., Editors (1980). *Studies in Subjective Probability*. Robert E. Krieger Publishing Co.
- Lad, F. (1996). *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. John Wiley.
- Laplace, P. S. (1812). *Theorie Analytique des Probabilities*.³ Paris: Courcier.
- Laplace, P. S. (1814). *Essai Philosophique sur les Probabilities*.⁴ Paris.
- Le Cam, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symposium on Math. Statist. And Prob.* **1**, 128-156. University of California Press. p. 308. ***** Please check where the p.308 come from? ***
- Mathews, Robert A. J. (1998). Facts versus factions: The use and abuse of subjectivity in scientific research. Cambridge, England: The European Science and Environment Forum, Working Paper 2/98, September, 1998.⁵
- Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.
- Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall.
- Porter, T. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press.
- Porter, T. (1996). Statistics, social science, and the culture of objectivity. *Oesterreichische Zeitschrift für Geschichtswissenschaften* **7**, 177-191.
- Press, S. J., and Tanur, J. M. (2001). *The Subjectivity of Scientists and the Bayesian Approach*. John Wiley and Sons.

³The second, third, and fourth editions appeared in 1814, 1818, and 1820, respectively. It is reprinted in *Oeuvres Completes de Laplace*, Vol. VII, 1847, Paris: Gauthier Villars.

⁴This book went through five editions (the fifth was in 1825) revised by Laplace. The sixth edition appeared in English translation by Dover Publications, New York, in 1951. While this philosophical essay appeared separately in 1814, it also appeared as a preface to his earlier work, *Theorie Analytique des Probabilities*.

⁵See <http://www.esef.org>, p.5.

第四篇 (此文的摘要和緒論曾用作習題)

Application of One Sided t -tests and a Generalized Experiment Wise Error Rate to High-Density Oligonucleotide Microarray Experiments: An Example Using Arabidopsis

Abstract: **Motivation:** A formidable challenge in the analysis of microarray data is the identification of those genes that exhibit differential expression. The objectives of this research were to examine the utility of simple ANOVA, one sided t tests, natural log transformation, and a generalized experiment wise error rate methodology for analysis of such experiments. As a test case, we analyzed a Affymetrix GeneChip microarray experiment designed to test for the effect of a CHD3 chromatin remodeling factor, PICKLE, and an inhibitor of the plant hormone gibberellin (GA), on the expression of 8256 *Arabidopsis thaliana* genes.

Results: The GFWER(k) is defined as the probability of rejecting k or more true null hypothesis at a given p level. Computing probabilities by GFWER(k) was shown to be simple to apply and, depending on the value of k , can greatly increase power. A k value as small as 2 or 3 was concluded to be adequate for large or small experiments respectively. A one sided t -test along with GFWER(2)=.05 identified 43 genes as exhibiting PICKLE-dependent expression. Expression of all 43 genes was re-examined by qRT-PCR, of which 36 (83.7%) were confirmed to exhibit PICKLE-dependent expression.

Key words: ***** Please add keywords ***

1. Introduction

The advent of inexpensive microarray technology has enabled individual laboratories to easily obtain a global perspective on the expression pattern of thousands of genes. This powerful technology has allowed investigators to diagnose early cancers (Kim, J. W. and Wang, X. W., 2003; Zhang *et al.*, 2003), discover genes that contribute to quantitative traits (Gu *et al.*, 2002), and detect coordinated gene regulation during pivotal developmental events such as embryogenesis and sexual maturation (Girke *et al.*, 2000; Lo *et al.*, 2003; Ruuska *et al.*, 2002).

The first generation microarrays were generally based on two dye methodologies. These cDNA microarray experiments involve hybridizing two mRNA samples, each of which has been converted into cDNA and labelled with its own fluorophore, on a single glass slide that has been spotted with 10,000-20,000 cDNA probes. In contrast, more recent high-density oligonucleotide microarrays, such as those offered by Affymetrix, provide direct information about the expression levels in an mRNA sample and can have a much higher density (Yang and Speed, 2002).

The majority of methodologies for microarray analysis have been developed for two dye spotted arrays (Kerr *et al.*, 2000; Kerr and Churchill, 2001; Lee *et al.*, 2003; Nguyen *et al.*, 2004, for review see Quackenbush, 2001 and Yang and Speed, 2002). Unfortunately these two-dye spotted arrays also pose other statistical issues, such as normalization to correct for dye bias. Furthermore if more than 2 treatments are used, it is not possible to compare all treatments on the same chip thus necessitating an Incomplete Block Design (IBD) type design (Kerr and Churchill, 2001). As such, special experimental designs, such as the reference and rotational design are needed for correct analysis (Kerr and Churchill, 2001; Quackenbush, 2001 and Yang and Speed, 2002).

In contrast, oligonucleotide microarrays use a single dye technology and pose some advantages, including a greatly increased density of genes and simplified experimental design because treatment effects are tested independently on each chip, eliminating the need for IBD designs. Nevertheless, statistical issues remain, such as normality of residuals, homogeneity of residual variance, correlation of errors within an array, and correlation of biological samples across arrays.

Mixed model methods for analysis of microarray experiment, proposed by Wolfinger *et al.* (2001), solves most of these issues (see Craig *et al.*, 2003 for review). However, the complexity of analysis dramatically increases with these advanced methods. Unfortunately, many of the current practitioners of microarray technology do not possess the mathematical expertise necessary to meaningfully employ these methods. On the other hand ANOVA is a tool that is easy to implement with methods common to most researchers. Kerr and Churchill (2001) conclude that “The analysis of variance (ANOVA) is a natural tool for studying data from experiments with multiple categorical factors”.

The first objective of this research was to examine the utility of simple ANOVA for analysis of replicated oligonucleotide microarrays experiments. The motivation was given eloquently by Kerr and Churchill (2001) who stated “An advantage of model based data analysis such as ANOVA is that a model helps the analyst explore the data. If one finds a model inadequate, discovering *why* it is inadequate can help the analyst identify sources of variation and bias.” A secondary objective of this study was to show how using a one sided *t*-test can be used to increase power. The final objective was to introduce an alternative method to increase power by accepting a base number of false positive with high probability.

The ANOVA is particularly suited to analyzing data from microarray experiments that employ a replicated factorial arrangement of treatments. An example of such an experimental design is one in which the investigator looks at gene expression in wild-type and mutant plants in the presence or absence of an added chemical. Many microarray studies incorporate this type of experimental design, e.g. the response of genes in nontumorigenic and tumorigenic tissues to different concentrations of toxic or therapeutic drugs (Lundquist *et al.*, 2002; Martinez *et al.*, 2002) or the response of genes from different tissues to estrogen or other hormones (Abe *et al.*, 2003; Faccioli *et al.*, 2002; Fujita *et al.*, 2003; Goda *et al.*, 2002). This design easily extends into any number of genotypes (or tissues) by any number of developmental time points (or biochemical exposures).

The primary biological objective of this research was to understand how a CHD3-chromatin remodeling factor, PICKLE, and a plant growth regulator, gibberellin (GA), regulate gene expression during germination of Arabidopsis seeds (Rider *et al.*, 2003). PICKLE is necessary for repression of embryonic traits in Arabidopsis (Ogas *et al.*, 1997). Expression of the embryonic state in pickle seedlings is inhibited by the plant growth regulator gibberellin (GA) and is enhanced by application of uniconazole-P, an inhibitor of GA biosynthesis (Izumi *et al.*, 1985; Ogas *et al.*, 1997). Specifically, gene expression was examined in wild-type and *pickle* seeds grown in the absence and presence of 10^{-8} M uniconazole-P. Thus the genotypes were ‘wild type’ vs. the *pickle* mutant, and biochemical exposure was to either 10^{-8} M uniconazole-P or no uniconazole-P during seed germination.

Our working hypothesis was that PICKLE functions during germination to repress genes that promote embryonic identity. In support of such a hypothesis, the transcript levels of two positive regulators of embryogenesis, LEAFY COTYLEDON1 (LEC1) and LEAFY COTYLEDON2 (LEC2) (Lotan *et al.*, 1998; Stone *et al.*, 2001), are elevated during germination of pickle seedlings (Ogas *et al.*, 1999; Rider *et al.*, 2003). Our interest was to find new genes that exhibited PICKLE-dependent expression, i.e. were up regulated. As such, we had a natural one sided test.

2. 21 Biological Methods

Seeds and tissues from the *Arabidopsis pickle-1* mutant (in a Columbia ecotype background) and wild-type Columbia were used for all investigations. Plants were grown as described previously (Ogas *et al.*, 1997; Rider *et al.*, 2003).

The Affymetrix GeneChip *Arabidopsis* Genome Array⁶ contained 8256 sets of oligos representing approximately 30% of the *Arabidopsis thaliana* transcriptome. A 2×2 factorial arrangement of treatments were examined. The first treatment was genotype (*pickle* mutant vs. wild type), the second treatment was uniconazole (applied vs. control), the treatment combinations were designated pkl, Upkl, wt and Uwt (*pickle* mutant untreated, *pickle* mutant treated with uniconazole-P, wild type untreated, and wild type treated with uniconazole-P) were each represented by four biological replicates ($n = 4$) for a total of 16 chips (Rider *et al.*, 2003).

3. Statistical Methods

3.1 The ANOVA, partitions, and transformations

The model for the completely randomized design (CRD) associated with the k -th spot (or gene) is $Y_{ij}^k = \mu + \tau_i^k + \epsilon_{(i)j}^k$ where Y_{ij}^k is the expression (or log transform) for the k -th gene, in the j -th replicate of the i -th treatment; μ^k is the overall mean; τ_i^k is the effect of the i -th treatment on that gene, and $\epsilon_{(i)j}^k$ is random residual. For maximum information treatment effects are further partitioned into main effects and interactions. The partitioning should be reduced to single degree of freedom tests by use of orthogonal contrasts. Because the ANOVA must be completed for each spot on the array, methods to automate the test are needed. To accomplish this goal, we use the well-known result that any single degree of F tests can equivalently be constructed as a t -test (Gill, 1978). A simple t -test for any contrast can be computed with the means procedure in SAS or in any standard spreadsheet, such as Excel. The t -test also offers the advantage of being able to test for a one sided alternative. In some experiments, as in this one, the researchers may only be interested in genes that are either up or down regulated, as a result, the power to detect those genes will be greatly increased.

For expression type data, the variance is usually correlated with the mean, violating a critical assumption for the ANOVA. For such data, transforming to logs will usually correct this problem. Interpretation of log transformed data also better meets the interest of the biologist as significant differences are interpreted as being significant ratios on a non-transformed basis, i.e. the difference between logs of numbers is the same as the log of a ratio. A log base 2 is interpreted as fold change, while base 10 is interpreted as orders of magnitude difference. Natural logs have not been widely used for array data but perhaps represents the most valid biological interpretation due to kinetics. A common rate equation in chemistry is where the rate of change in product (∂Y) per unit of time (∂t) is proportional (c) to the product (Y), thus $\partial Y = cY \partial t$. The solution to this differential equation is $Y = ce^t$. Therefore by taking natural logs, the expression is linearized into a rate equation, $\ln(Y) = \ln c + t$. If t is constant across biological replications, then variation in expression is due to linear differences in the rate constant c , the gene regulatory factor. Differences due to treatments are then interpreted as linear differences in gene regulatory factors (rate constants).

The vast majority of array data will require such a transformation, however, curiously these data better met the assumption when non-transformed. To check this assumption for any data, compute the within gene variance for each gene (the residual error variance in the ANOVA), then plot that against the average expression level for that gene. Any slope significantly different from

⁶part no. 510429, Affymetrix, Santa Clara, CA

zero (a zero slope is parallel to the x axis) indicates that the data require a transformation before the analysis proceeds.

For a given gene, because each treatment combination was randomized onto each of 4 biological replicates, the experiment as detailed above is a 2×2 factorial arrangement of treatments in a completely randomized design (CRD). The ANOVA for this design with treatment effects partitioned is given in Table 1.

Table 1: ANOVA table with partitions.

Source of Variation	Degrees of Freedom	Mean Square
Treatments	$t - 1$	MS(T)
Genotypes (C_1)	1	MS(C_1)
Inhibitor (C_2)	1	MS(C_2)
Interaction of Genotype x Inhibitor (C_3)	1	MS(C_3)
Within Error	$t(r - 1)$	MS(E)

The mean squares for the partitions can be found using the following formula along with the contrast coefficients given in Table 2

Table 2: Coefficients for partitions of treatment effects.

Treatment	Treatment Combination		Contrast Coefficients (C_{mj})		
	Genotype	Inhibitor	C_{1j}	C_{2j}	C_{3j}
1- pkl	<i>pickle</i>	None	1	1	1
2- Upkl	<i>pickle</i>	Uniconazole	1	-1	-1
3- wt	wild type	None	-1	1	-1
4- Uwt	wild type	Uniconazole	-1	-1	1

$$MS(C_m) = r \left(\sum_j C_{mj} \bar{Y}_{ij} \right)^2 I \left(\sum_j C_{mj}^2 \right). \quad (3.1)$$

The F test, which is distributed as F with 1 and $t(r - 1)$ degrees of freedom, is then computed as the ratio of $F = MS(C_m)/MS(E)$. This test is equivalently computed as t

$$T_m = \frac{\sum_j C_{mj} \bar{Y}_{ij}}{\sqrt{\frac{1}{r} MS(E) \sum_j C_{mj}^2}} \quad (3.2)$$

which has $t(r - 1)$ degrees of freedom. From these formula it is easy to verify that the calculated value $F = t^2$ and from tables one can verify corresponding critical values, i.e. $F_{1,t(r-1)} = (t_{t(r-1)})^2$.

However, when calculated as a t -test the sign of the contrast is preserved, thus allowing a one tailed test. This approach will extend to any contrast for any number of treatments, provided the sum of the coefficients for that contrast is zero. To be orthogonal with other contrasts the sum of the cross products must also sum to zero.

For this analysis, our hypothesis was that one or more genes existed for which the expression level was elevated in pickle mutants, regardless of uniconazole treatment. This hypothesis was based on an expression pattern similar to that of LEC1 and LEC2. Thus the primary contrast of interest was the main effect of genotype (C_1). Because we were only looking for a similar pattern (up regulation), the power to detect up regulated genes increased. Use of prior information to increase power is more cost effective than increasing the number of biological replicates. In other experiments additional contrasts may be of equal or greater importance, this may be particularly true of the interaction of genotypes with uniconazole treatment (C_3), which test the hypothesis that application of uniconazole has a different effect on one genotype than the other.

The critical value of t depends on a number of factors, including one- vs. two-sided alternatives, degrees of freedom (df) for estimation of error variance, and acceptable type I error rates. Choosing an acceptable Type I error rate is discussed in the next section.

3.2 Generalized experiment wise error rate (GFWER(k))

Experimenters have long recognized that if a comparison wise type I error rate (CWER) is used across a great number of tests, a large proportion of declared significant differences would be false. For example analysis of array data involves thousands of comparisons, consequently, if a per comparison error rate of 0.05 were used for our analysis, more than 413 of the 8256 tests would be expected to be declared significant by chance alone. The most widely used approaches to control Type I errors in multiple tests is based on controlling the family wise Type I error rate (FWER) (Fernando *et al.*, 2004). The FWER is the probability of rejecting one or more true null hypotheses, i.e. the probability of accepting one or more false positives. A common method for controlling the FWER is the Bonferroni or Sidak (1967) adjustments.

However, the FWER with those adjustments is too conservative if the cost of false negatives is high relative to the cost of false positives, i.e. they sacrifice power to avoid accepting false positives. Methods have been developed to address this issue by allowing for some false positives among those declared significant, such as the false discovery rate (FDR, Benjamini and Hochberg, 1995; Reiner. *et al.*, 2003; see Nguyen, 2005 for general discussion on this issue). Alternatives to the FDR have since been proposed that take into account the expected number of false null hypothesis and other modifications (see Fernando *et al.*, 2004 for review). However, all methods used to estimate an FDR make assumptions about the distribution of truly expressed genes. As a result, the FDR will either be too liberal or conservative.

Here we present an alternative that does not attempt to establish an FDR. Rather the method is an extension of the FWER methodology to allow for a higher family wise error rate. The development is as follows: Assume a strictly null distribution from which N independent test statistics are computed, from which N independent decisions are made at the same critical threshold level. The probability that any one decision is incorrect is p . An incorrect decision is defined as rejecting a true null hypothesis. With multiple tests, the probability of exactly m incorrect and $N - m$ correct decisions is

$$P(m = \text{incorrect} | N = \text{decisions}) = \binom{N}{m} p^m (1 - p)^{N-m} \quad (3.3)$$

The usual FWER = $\xi(1)$ is the probability of rejecting 1 or more true null hypotheses found as:

$$\xi(1) = \sum_{m=1}^N p^m (1-p)^{N-m} \quad (3.4)$$

or equivalently, 1 minus the probability of no incorrect decisions,

$$\xi(1) = 1 - (1-p)^N \quad (3.5)$$

which is Sidak's (1967) equation. The value of p per comparison (CWER) is found such that the $\xi(1)$ is achieved, i.e.

$$p = 1 - e^{\{\ln\{1-\xi(1)\}/N\}} \quad (3.6)$$

Stated in the reverse, there is a 1-FWER probability of *no* incorrect decisions among the N decisions made, i.e.

$$\omega(k) = \sum_{m=1}^{k-1} \binom{N}{m} p^m (1-p)^{N-m} \quad (3.7)$$

A generalization of this procedure is to divide the total probability of making Type I errors into parts associated with how many errors are likely to be made at a given probability. Among the N decisions made, define $\xi(k)$ as the probability of rejecting k or more true null hypotheses and $\omega(k)$ as the probability of rejecting fewer than k true null hypotheses, $\xi(k) + \omega(k) = 1$, where

$$\xi(k) = \sum_{m=k}^N p^m (1-p)^{N-m} \quad (3.8)$$

$$\omega(k) = \sum_{m=0}^{k-1} p^m (1-p)^{N-m} \quad (3.9)$$

If for a given k , the value for $\xi(k)$ is set to a small value, then among those tests declared significant, one accepts that there will be a high probability of $k-1$ false positives plus a low probability of k or more false positives. Therefore, a new type of error rate is defined as GFWER(k), which is strictly the probability of making k or more incorrect decisions at a given level of p , and ignores the probability of less than k Type I errors. The latter type of errors are considered acceptable in order to gain power and decrease the Type II error rate. For a more general development of the generalized family wise error rate see van der Laan (2004). The GFWER(k) cannot be solved for directly, but solutions can be found by iteration. SAS source code used to compute adjusted p values for any $\xi(k)$ and N is given at our web site. However, Equations (3.8) and (3.9) can also be approximated by the normal as follows: If X is binomial with n trials and probability of success p , then

$$P[X > r] \approx \Phi \left(\frac{r - np}{\sqrt{np(1-p)}} \right),$$

where Φ is the cumulative distribution of standard normal distribution.

Tables 3 and 4 give p values for, respectively, a one- and two-tailed alternative, and $\xi(k) = .05$. Associated critical values of t are given in Tables 5 and 6 for experiments with 6 and 60 df for estimating error variance. Note that for all k values, the critical value of t for a one-sided test

is between 12 and 13% smaller, with corresponding increases in power. Table 3 shows that by allowing for 2 or more false positives in a 1 sided t-test increases the adjusted p value by 6.5 times, and thereby also increasing the power of the test. Results presented in Tables 3 show that for the range of N examined (i.e. $N > 1000$) the ratios of p values for $\text{GFWER}(k)$ to that of $\text{GFWER}(1)$ are independent of N . Thus, for such chips, once the Sidak p values are found, $\text{GFWER}(k)$ can be found by multiplication using the constants given in the table.

Table 3: Adjusted p values for $k = 1$ to 5, chips of size 1,000 and 50,000 and a one-tailed $\text{GFWER}(k)=5\%$.

k	Number of Tests			
	1,000		50,000	
	$p\text{-value} \times 10^6$	Ratio*	$p\text{-value} \times 10^6$	Ratio*
1	51.29		1.025	
2	335.02	6.5	6.70	6.5
3	783.41	15.3	15.66	15.3
4	1320.01	25.7	26.38	25.7
5	1913.31	37.3	38.23	37.3

* Ratio of p -values to that of $\text{GFWER}(1)$

Table 4: Adjusted p -values for $k = 1$ to 5, chips of size 1,000 and 50,000 and a two-tailed $\text{GFWER}(k)=5\%$.

k	Number of Tests	
	1,000	50,000
	$p\text{-value} \times 10^6$	$p\text{-value} \times 10^6$
1	25.63	.521
2	167.5	3.35
3	391.7	7.83
4	660.0	13.19
5	956.65	19.15

An important issue is what value of k should one use. The value of k should be set as small as possible without sacrificing too much power. For an experiment of a given size, the rate at which power increases is dependent on the critical value of t . Examination of Tables 5 and 6 shows that the greatest decrease in the critical value of t , with either large or small experiment, comes from increasing k from 1 to 2. For large experiments increasing k beyond 2, or for small increasing k beyond 3, brings about much smaller incremental decreases in t . From these results, some general guidelines can be deduced for choice of k . Regardless of the number of spots on a chip, a k value of 2 or 3 should be adequate for large and small experiments respectively.

Table 5: The six genes for which the qRT-PCR assay detected no expression in untreated wild type seed. Transcripts were detected in untreated *pickle* seeds. Transcripts were also detected for both wild type and *pickle* seeds (Uwt and Upkl) germinated in the presence of uniconazole-*p*, thus permitting calculation of Upickle fold change relative to Uwt. The mean values from the arrays are included for illustration. #Pr is the number of times Affymetrix Microarray Suite software (v. 5.0) labeled a gene 'present' for the 16 gene chips used for this investigation.

AGI Code	#Pr	Mean values (4 chips)				qRT-PCR fold change				Putative ID/function
		wt	pickle	Uwt	Upickle	wt	pickle	Uwt	Upickle	
At3g16410	16	4535	14225	3630	18074	—	+	1	204.3	Jacalin type lectin
At4g27140	15	1000	2365	417	3214	—	+	1	8.95	2S1 seed storage protein
At1g67330	4	170	701	97	1255	—	+	1	3.75	uncharacterized
At5g13930	16	18459	37623	18471	46304	—	+	1	1.75	TT4/chalcone synthase
At3g23220	16	1611	2606	1364	2560	—	+	1	1.43	ERF1/transcription factor
At1g09750	16	2005	3200	2178	5193	—	+	1	0.58	nucleoid-like protein

Table 6: Presence of uniconazole-*p* increases derepression of *PICKLE*-dependent genes in *pickle* seedlings.

AGI Code	qPCR Ratios			Putative Function
	pkl/Wt	Upk/Wt	Upk/pkl	
At5g01600	2.90	8.82	3.04	maturation
At3g16420	4.75	11.41	2.40	defense
At1g73190	1.79	2.96	1.65	maturation
At2g28790	1.77	2.90	1.64	
At1g20620	2.33	3.51	1.50	maturation
At5g54740	2.97	4.10	1.43	
At4g19810	2.43	3.33	1.37	
At3g52500	5.55	7.55	1.36	
At3g16430	1.55	2.08	1.34	defense
At1g05510	1.88	2.53	1.34	
At3g16460	4.64	5.22	1.13	defense
At4g08685	2.61	2.78	1.06	
At2g35810	4.24	4.24	1.00	
At2g19590	2.50	2.24	0.90	
At4g37410	2.38	1.95	0.82	
At5g12030	5.36	2.68	0.50	desiccation

4. Biological Verification: QRT-PCR analysis

Those genes found significant with ANOVA were re-analyzed using qRT-PCR to compare results. The qRT-PCR method, while more precise than the chip analysis, is still subject to error. The method is based on PCR amplification of mRNA in the sample until a pre-determined threshold is obtained. Because the amplification is a doubling with each cycle, the accuracy of the method is questionable if there exists less than a 2 fold difference in mRNA between the two treatments. qRT-PCR is also subject to biological variability between samples and should therefore also be replicated and treated to statistical analysis. However, replicated qRT-PCR analysis for

each gene would be extremely expensive and time consuming. Therefore within the limitations of this experiment, and recognizing those limitations, we defined confirmation of *PICKLE*-dependent expression as a two-fold or greater increase in expression level of a given gene in *pickle* versus wild-type seed when grown in either the absence or presence of uniconazole-P. qRT-PCR was used to compare transcript levels in *pickle* versus wild-type seed grown in the absence of uniconazole-P as well as transcript levels in *pickle* versus wild-type seed grown in the presence of uniconazole-P.

Quantitative RT-PCR was performed on an ABI sequence detection system using RNA from one of the biological replicates previously generated (Rider *et al.*, 2003). Oligonucleotide primer sequences and primer concentrations used are listed in supplementary Table 2S available at the web site.

5. Results and Discussion

5.1 Statistical issues

For this experiment, we used $\xi(2) = .05$. Allowing for one false positive raised the adjusted p value from 6.21×10^{-6} to 4.1×10^{-5} and correspondingly increased the power of the test. The ANOVA method selected 43 genes, less than one of which was expected to be a false positive based on the experimentwise selection criteria that we employed ($8256 \times 4.1 \times 10^{-5} = .33$). Our qRT-PCR analysis supported 36 of the 43 genes (Figure 1). A surprising result of this study was that qRT-PCR did not detect transcripts in wild-type seeds for 6 of the 43 genes identified as having expression differences based on analysis of the array data (Table 5). Although this observation is consistent with the hypothesis that *PICKLE* represses expression of these genes in wild-type seeds to facilitate the developmental transition from embryo to seedling, the array expression values did not suggest absence of transcripts in wild-type seeds.

figure 1 about here

There are at least two possible explanations for the elevated number of observed false positives. Affymetrix constructed this GeneChip when the sequence of the Arabidopsis genome was only partially completed. Inflated expression values for some oligos may have arisen from cross hybridization to unintended targets. In fact, two of the false positives were false because qRT-PCR detected no expression in germinating seeds under any condition. Alternatively, as previously discussed, the discrepancy may be due to different criteria used to determine success for each method. The qRT-PCR data should only be viewed as supporting evidence, not confirmatory.

5.2 Biological Inferences

PICKLE is necessary to repress expression of embryonic traits in Arabidopsis seedlings. Previous analysis of genes that exhibit *PICKLE*-dependent repression identified genes associated with various stages of seed development. ANOVA identified genes associated with seed development, including 2S albumin genes, HSP17.6, and several lectin-like genes (Guerche *et al.*, 1990; Lenman *et al.*, 1993; Ruuska *et al.*, 2002; Sun *et al.*, 2001). In all, 10 of the genes (28%) identified and confirmed by qRT-PCR analysis were associated with embryo development or exhibit sequence similarity to genes involved in embryo development (Table 1S, available at the web site). Additional studies will be necessary to determine if the other 26 genes that showed *PICKLE*-dependent expression in the germinating seed are also involved in some aspect of embryo development. Previous expression analysis did not suggest a specific role for uniconazole-P in increasing penetrance

of the pickle root phenotype in *pickle* seedlings (Rider *et al.*, 2003). This analysis revealed that the extent of derepression of many of the genes that exhibit *PICKLE*-dependent repression is enhanced by the presence of uniconazole-P (Figure 1, black bars versus white bars). The magnitude of this enhancement, however, was often due in large part to the fact that the presence of uniconazole-P resulted in decreased expression of the gene relative to wild-type seed imbibed in the absence of uniconazole-P (data not shown).

In order to examine the effect of combining the *pickle* mutation with exposure to uniconazole-P, we compared the fold change values of genes in *pickle* versus wt seedlings (pkl/wt) and the fold change values of genes in *pickle* treated with uniconazole-P versus wt seedlings (Upkl/wt) as determined by qRT-PCR (Table 6). In order to make this analysis comparable to previous analysis of the dataset, either the ratio pkl/wt or the ratio Upkl/wt or both had to be ≥ 2 for a gene to be included in this analysis. Genes for which a transcript was not detected in wild-type seedlings were excluded from this analysis. Sixteen genes identified with ANOVA met these expression criteria. We found that the presence of uniconazole-P did increase expression of many of these genes in *pickle* seedlings; the transcript level of 10 genes increased 33% or more when *pickle* seeds were imbibed in the presence of uniconazole-P. In contrast, a previous analysis of the same array data identified no genes for which the corresponding transcript was increased by treatment of *pickle* seeds with uniconazole-P (Rider *et al.*, 2003).

Uniconazole-P increases the probability that primary roots of the *pickle* mutant will express embryonic differentiation traits (Ogas *et al.*, 1997). Genes associated with seed development exhibit elevated expression in *pickle* seedlings, suggesting that the expression of these genes contributes to the ability of *pickle* seedling to express embryonic traits after germination (Rider *et al.*, 2003). The discovery that the presence of uniconazole-P enhances the expression of 10 genes in *pickle* seedlings, 5 of which (50%) are involved in seed development or exhibit sequence similarity to genes involved in seed development, suggests for the first time that the increased penetrance of embryonic traits in *pickle* seedlings treated with uniconazole-P may be mediated in part through changes in gene expression. Specifically, our results are consistent with the hypothesis that GA acts in concert with *PICKLE* during germination to repress expression of genes that promote embryonic traits. Further characterization of the genes identified here may facilitate subsequent genetic and biochemical analysis of the GA signal transduction pathway that mediates this response.

5.3 Utility

We have shown that a simple ANOVA method can identify a manageable number of candidate genes for differential expression from a gene expression array, most of which were real. Although we only applied the approach to an experiment that incorporated a simple class-by-treatment design, it is applicable to any full factorial design and is computationally straightforward. Previous analysis of the array data employed a modified fold change (MFC) approach (Rider *et al.*, 2003) and failed to detect many of the genes identified by ANOVA. In addition, our current analysis demonstrates that treatment of *pickle* seedlings with uniconazole-P enhances the derepression of *PICKLE*-dependent genes during germination. These results reinforce the power of ANOVA versus a method that emphasizes fold-change.

The practical utility of the GFWER(k) method is derived from allowing the user to influence the number of genes identified by selecting the appropriate value for k , the number of false positives allowed above the threshold significance level. A critical question is what value of k will result in the greatest increase in power with the lowest number of Type I errors. A simple power analysis showed that regardless of the number of spots on a chip, a k value of 2 or 3 should be adequate for large and small experiments respectively. Although the GFWER(k) and FDR are closely related and greatly increase the power of the experiment by relaxing the Type 1 error rate, the application of the GFWER(k) does not attempt to project an FDR, rather, we only set the maximum number of false

positives under the null hypothesis. Calculations for an exact FDR would require knowledge of 1) the number of truly expressed genes, 2) the signal to noise ratio, and 3) their distribution. Without knowing these factors, the FDR as calculated by any of the current methods is an approximation. As a result, the GFWER(k) may be more or less conservative than FDR methods, depending on the particular experiment. However, the GFWER(k) is constant and independent of the experiment, which in itself is appealing. This gives rise to another interesting difference between the methods. The expected number of false positives can be determined *a priori* with the GFWER(k) because the rate is independent of the data, whereas with the FDR (and newer methods as reviewed by Fernando *et al.*, 2004) calculations are dependent on the data and one has to wait until the list is generated to determine what the expected number of false positives will be. This difference could be critical in the planning stage of an experiment.

Acknowledgements

We thank Guilherme Rosa for helpful comments and discussion during preparation of the manuscript, Jim Henderson for helping to generate the RNA used for this study and Howard Edenberg for chip hybridization and calculation of expression values. **Funding:** Partial Funding for this research was provided by: JO National Institutes of Health (R01GM059770-01A1 and 5R01GM059770-02); JRS The Indiana 21-st Century Research and Development Fund and Purdue Research Foundation; SDR BASF ; WM USDA/NRI 9803430 Animal Genetic Mechanisms. This is journal paper number 17605 of the Purdue University Agricultural Experiment Station. **Supplemental tables** used in the analysis are available at: http://www.biochem.purdue.edu/research/ogas_lab/arrays/JPmethod/.

References

- Abe, H., Urao, T., Ito, T., Seki, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2003). Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **15**, 63-78.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate — A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289-300.
- Black M. A., and Doerge R. W. (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics* **18**, 1609-1616.
- Craig B. A., Black M. A., and Doerge R. W. (2003). Gene expression data: The technology and statistical analysis. *J. Agricultural Biological and Environmental Statistics* **8**, 1-28.
- Faccioli, P., Lagonigro, M., Cecco, L., de, Stanca, A., Alberici, R., Terzi, V., and de Cecco, L. (2002). Analysis of differential expression of barley ESTs during cold acclimatization using microarray technology. *Plant Biol.* **4**, 630-639.
- Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., and Soller M. (2004). Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**, 611-619.
- Fujita, N., Jaye, D. L., Kajita, M., Geigerman, C., Moreno, C. S., and Wade, P. A. (2003). MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer. *Cell* **113**, 207-219.
- Gill, J. L. (1978). *Design and Analysis of Experiments in the Animal and Medical Sciences. Volume 1.* The Iowa State University Press.

- Girke, T., Todd, J., Ruuska, S., White, J., Benning, C., and Ohlrogge, J. (2000). Microarray analysis of developing Arabidopsis seeds. *Plant Physiol* **124**, 1570-1581.
- Goda, H., Shimada, Y., Asami, T., Fujioka, S., and Yoshida, S. (2002). Microarray analysis of brassinosteroid-regulated genes in arabidopsis. *Plant Physiol* **130**, 1319-1334.
- Gu, C. C., Rao, D. C., Stormo, G., Hicks, C., T., and Province, M. A. (2002). Role of gene expression microarray analysis in finding complex disease. *Genes. Genet Epidemiol* **23**, 37-56.
- Guerche, P., Tire, C., De Sa, F. G., De Clercq, A., Van Montagu, M., and Krebbers, E. (1990). Differential expression of the arabidopsis 2S albumin genes and the effect of increasing gene family Size. *Plant Cell* **2**, 469-478.
- Izumi, K., Kamiya, Y., Sakurai, A., Oshio, H., and Takahashi, N. (1985). Studies of sites of action of a new plant growth retardant (E)-1-(4-chlorophenyl)-4,4-dimethyl-2-(1,2,4-triazol-1-yl)-1-penten-3-ol (S-3307) and comparative effects of its stereoisomers in a cell-free system from Cucurbita maxima. *Plant Cell Physiol* **26**, 821-827.
- Kerr M. K., and Churchill G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**, 123-128.
- Kerr M. K., Martin M., and Churchill G. A. (2000). Analysis of variance for gene expression microarray data. *J. Computational Biology* **7**, 819-837.
- Kim, J. W., and Wang, X. W. (2003). Gene expression profiling of preneoplastic liver disease and liver cancer: A new era for improved early detection and treatment of these deadly diseases? *Carcinogenesis* **24**, 363-369.
- Lee, M. T., Whitmore, G. A., Yukhananov, R. Y. (2003). Analysis of unbalanced microarray data. *Journal of Data Science* **1**, 103-121.
- Lenman, M., Falk, A., Rodin, J., Hoglund, A. S., Ek, B., and Rask, L. (1993). Differential expression of myrosinase gene families. *Plant Physiol* **103**, 703-711.
- Lo, J., Lee, S., Xu, M., Liu, F., Ruan, H., Eun, A., He, Y., Ma, W., Wang, W., Wen, Z., and Peng, J. (2003). 15,000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Research* **13**, 455-466.
- Lotan, T., Ohto, M., Yee, K. M., West, M. A., Lo, R., Kwong, R. W., Yamagishi, K., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (1998). Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. *Cell* **93**, 1195-1205.
- Lundquist, H., Oredsson, S., and Akesson, B. (2002). Effect of quercetin on gene expression in human cells as measured by microarrays — a pilot study. Paper presented at: Health promoting compounds in vegetables and fruit. Proceedings of a workshop in Karrebaksminde, Denmark, 6 8 November, 2002. DIAS Report, Horticulture. 2002, No.29, 77 80; 10 ref. (Danmarks JordbrugsForskning; Tjele; Denmark).
- Martinez, J. M., Afshari, C. A., Bushel, P. R., Masuda, A., Takahashi, T., and Walker, N. J. (2002). Differential toxicogenomic responses to 2,3,7,8-tetrachlorodibenzo-p-dioxin in malignant and nonmalignant human airway epithelial cells. *Toxicol Sci.* **69**, 409-423.
- Nguyen, D. V., Wang, N. and Carroll, R. J. (2004). Evaluation of missing value estimation for microarray data. *Journal of Data Science* **2**, 347-370.
- Nguyen, D. V. (2005). A unified computational framework to compare direct and sequential false discovery rate algorithms for exploratory DNA microarray studies. *J. Data Science* (In Press).
- Nitin, J., Thatte, j. Braciale, T., Ley, K., O'Connell, M. and Lee. J. K. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **19**, 1945-1951.
- Ogas, J., Cheng, J. C., Sung, Z. R., and Somerville, C. (1997). Cellular differentiation regulated by gibberellin in the arabidopsis thaliana pickle mutant. *Science* **277**, 91-94.

- Ogas, J., Kaufmann, S., Henderson, J., and Somerville, C. (1999). PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in arabidopsis. *Proc. Natl. Acad. Sci. USA* **96**, 13839-13844.
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics* **2**, 418-429.
- Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-375.
- Rider, S. D., Henderson, J. T., Jerome, R. E., Edenberg, H. J., Romero-Severson, J., and Ogas, J. (2003). Coordinate repression of regulators of embryonic identity by PICKLE during germination in arabidopsis. *Plant J.* **35**, 33-43.
- Ruuska, S. A., Girke, T., Benning, C., and Ohlrogge, J. B. (2002). Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell* **14**, 1191-1206.
- Sidak, Z. (1967). Rectangular confidence regions for means of multivariate normal distributions. *J. Amer. Statist. Asso.* **62**, 626-*****
- Sokal, R. R., and Rohlf, F. J. (1995). *Biometry, Third edition*. W. H. Freeman and Company.
- Stone, S. L., Kwong, L. W., Yee, K. M., Pelletier, J., Lepiniec, L., Fischer, R. L., Goldberg, R. B., and Harada, J. J. (2001). LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. *Proc. Natl. Acad. Sci. USA* **98**, 11806-11811.
- Storey, J. D. and Tibshiran, R. (2003a). Statistical significance for genome wide studies. *Proc. Natl. Acad. Sci, USA* **100**, 9440-9445.
- Storey, J. D. and Tibshirani, R. (2003b). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software* (Edited by G. Parmigiani, E. S. Garrett, R. A. Irizarry, S.L. Zeger). Springer.
- Sun, W., Bernard, C., van de Cotte, B., Van Montagu, M., and Verbruggen, N. (2001). At-HSP17.6A, encoding a small heat-shock protein in Arabidopsis, can enhance osmotolerance upon overexpression. *Plant J.* **27**, 407-415.
- van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*⁷
- Wolfinger R. D., Gibson G., Wolfinger E. D., Bennett L., Hamadeh H., Bushel P., Afshari C., Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Computational Biology* **8**, 625-637.
- Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics* **3**, 579-588.
- Zhang, H., Yu, C. Y., and Singer, B. (2003). Cell and tumor classification using gene expression data: Construction of forests. *Proc. Natl. Acad. Sci. USA* **100**, 4168-4172.

September 4, 2006

⁷See <http://www.bepress.com/sagmb/vol3/iss1/art14/>

第九部分：讀一篇關於 EM 的論文

這是一篇「集大成」型的論文。對於統計推論——尤其是古典的 parametric inference——基本上可將問題簡化為：將 likelihood function $L(\phi|x)$ 找出來。因為一旦找了出來，計算 MLE $\hat{\phi}$ 只是一個優化型的技術問題。有了電腦、有了軟體、只要又肯花錢，只要未知參數 ϕ 的維數不高，這工作總是可以做的。

對有些問題來說，求出 likelihood function 就難，計算 MLE 當然就更難了。若是我們將可以容易寫出 likelihood function 的數據叫做「完整數據¹ (complete data)」，那麼較難寫出 likelihood function 的數據，是不是可以經由完整數據的 likelihood function 來寫？

之所以樣的想法，多因為不完整數據有時是因為我們嘗試求得完整數據失敗而得（例如問卷中的 non-response, partial answer 等）。假如有一組數據，它可以想像成由某一組完整數據「丟掉一部分訊息」而來，那麼我們可不可以利用原來的 likelihood function 來求出相對應的、關於這組不完整數據的 MLE 來呢？

這篇文章不管求出來的 MLE 對於統計推論是好是不好，只論求 MLE 的方法。它的賣點在：對於所謂的「不完整數據」，作者有明確而廣泛的看法，而且可以舉出很多例子來——因此對應用者來說，是非常有吸引力的。同時，作者將「求不完整數據的 MLE」的基本技術，有點像「分解動作」那樣地歸納成兩個步驟，分別叫做 E 步 (E step) 和 M 步 (M step)。其中的 M 步，對於 exponential family 而言，還含有「可儘量利用關於完全數據求 MLE 的軟體」的能力。換句話說，對很多不完整數據的問題而言，你可以在現有的軟體上只增加一個計算迴路 (loop)，就可以算出想要的 MLE。

讀我的註脚需參考原始論文。見

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), *Journal of the Royal Statistical Society, Ser. B* **39**, 1-22.

以下為我的註脚。

p.1, summary. 第一句話便說明了本文的全部內容，是名家手筆。用字不多，但恰到好處。第二句話直指技術面，而 monotone 是主要的技術因素（因此收斂是必然的）。由最後一句，便引出本文的一大堆例子來。

這些例子，有很多是被前人考慮過的。有些有 EM 的想法，有些沒有（或者那些作者沒有明白說出）。但將它們都看成一類而總體都能解，並有實際的為法來解，當然是重要的結果。

這是寫得極好的論文摘要，用字不多（也沒有用任何困難的字²），但全文的內容都被

¹這是本文的簡單看法。因為要從完整才能定義不完整。世上當然有我們所直觀了解的完整數據，但又寫不出 likelihood function 的。

²一般的統計論文，除了（專有）名詞外。大多不用難字——文章好不是用字難或不難，而是用得恰當與否。

涵蓋。

p.1, introduction, line 1. 這一句和摘要是一意思, 但多了 *iterative* 一詞, 講法就不大一樣。有些作者喜歡將緒論的內容差不多照抄到摘要 —— 這是壞習慣, 因為明擺著不肯用心, 容易引起 referee 反感。

p.1, introduction, line 4. 一般論文極少自誇。此處用 *remarkable* 已有一點算是自誇³。這類事大教授可以做, 因為他們的身份和品味擺在那兒。新手以不做為宜。但 “specify, generality, wide range of examples” 是用得好的。

p.1, introduction, line 6. 在此先做 *exponential family*, 是因為大家在數理統計裡一開始就會讀到, 是耳熟能詳的。這是統計理論上最好用的模型: 所有的好方法, 在 *exponential family* 上就一定能用而且會簡化到有直觀意義。當然, 若是用不上, 便強烈地暗示「所提的方法有問題」。

p.1, introduction, line 9 – (1.1). 這一段定義何謂 *incomplete data*: 從某一組 *complete data* 經簡化、遺漏...而來。但真的怎樣來? 用嘴巴是說不清楚的, 必須用數學式子, 因為後面要做證明, 而空口白話是無法證明的。

要做研究就得把何謂 *incomplete data* 說明白。我的意見是: 公式 (1.1) 是本文的主關鍵之一: 用「積分」來說明白。對 *complete data* 的 *likelihood function* 積掉那些看不到的部分, 便得到 *incomplete data* 的 *likelihood function*。以 Dempster 和 Rubin 的功力 (Laird 是 75 年的博士, 那時還是小教授), 在有了 (1.1) 之後, 本文其它的部分, 應都可手到拈來。

但能看出 (1.1) 可不簡單。當然, 看出來後再舉例子就容易了。例如若有 $X_1, \dots, X_{100}$ 算是 *complete data*, 它的 *likelihood function* 是 $f(x_1, \dots, x_{100}|\phi)$ 。若我們丟掉了 x_{100} , 則不完整的數據就是 $X_1, \dots, X_{99}$, 而我們可以用積分的方法找到這二者者的關係:

$$f(x_1, \dots, x_{99}|\phi) = \int f(x_1, \dots, x_{99}, x_{100}|\phi) dx_{100}$$

而這就是 (1.1)。

這一段明確地交待了 *incomplete data* 和 *complete data* 間的 *likelihood functions*, 應如何連接。以後的工作, 只是將這個現象用證明及例子講得清楚。

p.1, introduction, line 20. 這一段只是說, 要「充分利用 $f(\mathbf{x}|\phi)$ 」。求 (1.1) 的極大有兩個辦法。(1) 直接把積分算出來, 再求極大, 但此法的條件是「能做出這個積分」; (2) 設法另尋蹊徑, 避免做 (一般幾乎做不出來的) 積分。

p.1, introduction, line 23. 說明 (1.1) 未必為 *unique*。當然, 我們可在 $X_1, \dots, X_{101}$ 中, 丟掉 X_{100}, X_{101} , 然後再得到 $X_1, \dots, X_{99}$ 。這也暗示了, 在實際的問題裡, 找到可用的 $f(\mathbf{x}|\phi)$ 可能並不容易。

p.1, introduction, line 25. 此處進入本文的標題, 謂之「點題」。注意到起名字是一個學問⁴。如 *bootstrap*, *jackknife*, *EM* 都取得不錯 —— 易記, 就會想到去用。

³但是論文又非得說自己的東西好。所以需要你能平實地說自己好。

⁴我以前工作的公司, 有一個部門叫 *Human Factor*, 專門替新產品取名字。

p.2. 此一頁只講一個例子。你可以直接用 (1.2) 來做, 這樣做要解一個三次方程式; 也可以用 EM 的想法, 此時不需要解任何方程式。即使最後到了 π^* (真正的收斂極限), 也只是解一個二次式。這一頁不難, 一般的博士生應可一步一步地跟著做。注意到作者原可自己設定一個例子, 但卻寧可用 Rao (1965)⁵ ——這也是一本經典教本。

p.2, line 4. 特別提起 genetic model, 雖然下文和遺傳一點關係也沒有。作者只是說這不是人造的不自然的例子, 是真有實驗的。

p.2, line 20. 提「只要八步」是故意的。

p.3, table 1. 最後一個 column 中的數字, 暗示的是: 收斂的速度, 是 exponential rate。但作者卻沒有講 —— 意在言外。

p.3, line 1 to end of section 1. 此一段敘述 EM 的歷史因緣和作者的看法。對 Hartley (1958) 這篇文章, 注意到作者雖說 “many times”, 但又加上 “in special circumstances” —— 故雖在給別人 credit, 卻不肯給足。

p.3, line 3 after table 1. 此處又將 EM 的廣度提了一層: 和 robust estimation 拉上關係。

p.3, lines 7-9. 注意到, 雖一直在提別人的工作, 但仍在說 “special examples”。這等於是說別人沒有進一步的理論 —— 一直都在說自己好, 卻不帶火氣。

p.3, line 10. 注意到 “always increase the likelihood”, 並說這是 key result: 對自己的結果, 定要說話。

p.3, line 14. Dempster 和 Rubin 都是 Bayesian 中的大家, 當然不會忘記 Bayesian application 了。

p.3, section 2. 先只講滿足 (3.1) 的指數族 (exponential family)。這一段和第一節的例子其實是一回事。只是前面是 special example, 而此處是一般的 exponential family。

p.3, section 2, lines 1-5. 這一段先自己說 “strong restriction”, 免得別人說。又說 section 3 的理論遠不止此, 但 exponential family 另有意義, 故值得另寫一節。

若在五零年代, 你可以分別做 binomial, Poisson, normal, gamma ...。但到了 1977, 就不能這樣地 elementary。

p.3, section 2, line 9, line 10, line 11. 注意用詞: “regular, convex, unique”。這些都是作者小心處。表示從數理統計做問題的嚴謹。又在 line 12 to p.4 line 1 中, 提出只做 natural parameter case, 一直到 p.4, line 6, 這些話都是為「小心」而說的, 只是表現作者的數理統計唸得很扎實。最後, 集合 $A_\phi = \{\mathbf{x} : f(\mathbf{x}|\phi) > 0\}$ 和 ϕ 無關, 是常用的假設, 此處要放上。因為若沒有它, MLE 的求法 (即使對於 complete data) 都可能困難, 本部的簡潔結論, 就得不出了。

p.4, lines 12, 14. 注意到, 提到 (2.3) 是用 “equations”, 因為它可看成一組方程式。你當然可以用 “equation”, 將之視作一個用 vector 來表的 equation 就是。但注意到需一

⁵此時若在一本爛書上找例子, 就不如自己來造。

致。即便是 likelihood equations 在此也用複數。

p.4, line 14. “familiar form” 此詞用得精準，因為 (2.3) 並不是 exactly a likelihood equation。

p.4, line 18. $\log$ 是不該用斜體字的。同理，在方程式裡要用 $\log, \sin, \tan, \dots$ 而不是 $\log, \sin, \tan, \dots$ 。但在 $\log x$ 裡的 x ，字體卻又有不同。將來在你送出的文稿中要注意這種小地方。

p.4, lines 18-22. 這一段繼續說，注意到作者一直在反覆地說明同一件事。甚至於 x 是整數，但 $E(x)$ 不是整數都交待明白——這些大教授是很小心的。

p.4, line 17. convexity 回應了 regular exponential family 的條件。主要是為了 (2.3) 的解是唯一——前面所埋的伏筆，儘管很不重要，最好都能關照——若不是為了要用 (2.3) 的解是唯一，何必在前面又說 convexity，又說 regular exponential family？

p.4, lines 11-28. 這一段若由國人來寫，多半是用數學式子一個一個地套下來。但統計學還有社會科學的味道，用太多公式，只是表示你只能推導式，卻不能解讀公式。所以，能用講的，就用講。在這裡最能表現你的深度。

p.4, line 29 – p.5, line 17. 這一段雖然客氣地叫做 digress to explain，但要點在 explain (若無此意，則不必 digress)。雖然全是推導，但皆是形式上的推導。最後得到的 (2.13)=0，卻正和 (2.2)=(2.3) at $\phi = \phi^\infty = \phi^*$ 一致。這是古典 EM 的關鍵。故需用種種角度，一提再提。

p.5, line 18. 注意 “in special cases by many examples”，仍然是給 credits，但不全給。作者絕口不提若無以前的 many examples，他們根本做不出來。國人的寫法太謙虛，不好，因為別人都不謙虛。

p.5, line 20. 提到 1966 的講義。告訴讀者「我們知道最早是誰做的」。在 1966，有人明明有好結果，但並不寫論文，只發一個講義而已。

p.5, line 20 – (3.16). 這一段是引自 Sundberg (1974)。Parenthetically 指「附加說明地」。這一段只是表示作者搞懂了這篇 1974 的文章。(2.16) 在 $k = 1$ 時恰好是 (2.13)。至於在 $k = 2$ 時恰好得到 covariance matrix 一事，本文雖有提到但未深究。後來 L. Thomas 因此做了一篇不錯的文章⁶。

p.5, line 36 – p.6, line 4. 這一段其實有點多餘。Curved exponential family 並不佔太多特殊位置，此處卻用它「不能得 (2.13)」的理由，來要改用新的 M-step。但此又為下面的一般 EM 所包括。就文論文，此一段沒有新東西且不特出，是可有可無的中間步。

p.6, line 5. 注意作者以前只用 “exponential family”，此就卻用 “exponential families”，何以故？以前只有一個 (指 regular exponential family)，但現在已有兩個 (regular exponential family and curved exponential family)。故用複數。

p.6, line 17. 此處又回到 exponential family，這相當於 checking。當你做了更廣的一步，常需要和已知的、知道是對的結果相互驗證。如果不合，就是某一就算錯了。

⁶L. Thomas (1982). *JRSSB* 44, 226-233.

p.6, line 22 to end of section 2. 這一段話把 Bayesian method extension 就全包括了。好好地改寫，可作一篇博士論文呢！

p.6, footnote. 注意到評審對於某一個字的字義的挑剔。英文用字，有時比中文要精確。其實中文也可以寫得精確，只是大家都有壞習慣了不去要求而已。國人的英文已不如洋人，如遣詞用句再下肯用心⁷，人家的地盤上，文章不易被接受，是明擺著的。慎之，慎之。

p.6, section 3. 此一節是技術性的：證明定理。這是國人學者的強項，但統計學的結果不在於數學的難度。這幾個定理都不能算難——只是說明 EM 在甚麼時候可以用而已。定理一當然不難，因為基本上，p.7 的 (3.6) 這個式子便已足夠。若自 (3.5) 來看，GEM 的定義，是能夠讓 $L(\phi) \uparrow$ 的算法而已（(3.3) 是 well known 的結果）。這是「看出定理較難，證明反而容易的定理。事實上，直到 p.8, line 3 都沒有太難。

p.8, line 12. 此一段說明前面證明的是甚麼。文中先說 $L(\phi^{(p)}) \uparrow$ ，再說 $\phi^{(p)} \rightarrow \phi^*$ ，最後還得去說 $\phi^{(p)} \rightarrow MLE$ 。

這類事一般都需要條件。所謂的好期刊，一般都會要求作者好好把條件列出來，好好證一證，並且希望你所用的條件不能再弱。

p.8, (3.15), (3.16). 符號 D^{10}, D^{20} 猛一看沒有定義。仔細想，其實 (3.15), (3.16) 就是定義。

p.8, line 32. 用 “an instance of a GEM” 是用語的小心。

p.8, line 32 to end of p.9. 這一段的證明，你看不看都無所謂。但對於「投稿一流期刊」而言，這樣的段子就非有不可。如果你要磨鍊自己手上功夫，不妨試作去補足所有的技術細節。這些一般在上課是不教（因為太 detail），但研究上你該都理清楚。這四個定理都在說甚麼？真正所需的條件為何？這類細節，將來你的論文做到有初步結果時都要一一補上⁸——對某些期刊而言，這才算 good quality work。

p.9, line 30. “can be easily verified ...” 雖是廢話，但也要說。因為否則這些定理的條件是否有模型能滿足都不知道⁹。但作者卻小心地說 “in many instances”，這暗示著 “not in all cases”，也是自我保護但又不負面的寫法。

這些都是號稱為了嚴謹而玩的文字遊戲。注意到真正的要點是 GEM 基本上是 by definition, it works。這些定理都是化妝後的結果：真正用 EM 去做問題的人，大概都不會去查條件滿足與否——除非做出來的結果自己都不信。

p.10, lines 1-25. 這一段說話，是為了表現學問而多寫的 junk。完全看不出作者是否真的證明過，還是憑經驗亂猜。但三個作者都是 Harvard 教授（Laird 那時是 assistant professor），應該不會弄虛作假。

p.10, lines 26-32. 這是為了回應 table 1, column 3。也可將 table 1, column 3 看作爲伏筆，現在才回應。

⁷將心比心，如果你評審一篇洋人用不通的中文稿，會讓他通過嗎？

⁸儘管可能放在附錄，或甚至要求你刪去，有細節而被要求刪去和無細節被認為不夠嚴謹而受拒，是兩回事。

⁹有時，為了證明定理就加條件，結果條件太多，以致「滿足該條件的情形」變成空集合。因此說明自己的模型的確存在。為了嚴謹也是重要的。

p.10, line 33 to p.11, end of section 3. 這裡在儘力向深處走, 但又語焉不詳。為何用到 second derivative 時, 便會 speed up convergence? (3.19) or (3.26) 說的是甚麼? 為何與收斂速度有關? 此類言語, 不容易說。因為說得不對時, 內行人便看得出來, 便成為畫虎類犬了。

自 line 45 起的一段話, 是回頭再做 literature review。前人的類似工作, 要一一承認。但作者仍然在說別人未竟全功 (如 “unusual special cases”, “without recognition of”, “do not focus on directly” 等。極力說前人雖有建樹但都缺臨門一脚。

最後, 還是回到 Bayesian, 和以前的話輝映。這是作者寫作用心處。

p.10, line 1 to p.11, end of section 3. 此段整體而言, 仍然是在說「我們對所舉的 EM 諸性質, 都儘量做了」。這是用來堵 referee 用的 (我們已非常努力了, 別多問細節了)。因為 $Q(\phi'|\phi)$ 這個函數得要算得出來才行, 其它都是空話, 因此文字方面有點枯燥, 就不足為奇了。

p.11, section 4, lines 1-9. 整個 section 4 都是在給例子。這裡說 “either has been or can be used”, 因此挑明了這些例子都不見得是作者的發明。這一節的主要目的是, 對於所舉的 incomplete data 的問題中, 將合理的 complete data 的形式想出來 (雖不一定 unique)。並導出 E 和 M 兩步。

p. 11, section 4.1.1 to p.21, end of section 4. 以下的各小節的寫法都頗類似 (但英文用字卻都有差別): 先介紹一個情形, 再將問題改裝成 incomplete data 的形式, 並指出相對應的 E 和 M 兩步, 應如何做。

總體而言, 這是一篇蠻囉嗦的文章。它的真實例容其實只有 (1.1) 和定理一, 其它的都是佐料。能舉出那麼多例子是作者們的本事。它表面不難, 但將那麼多以前的東西整理得只用 E 和 M 兩步就弄明白了, 是不簡單的。

更要緊的, 是這個文章有市場。它的架勢是: 幾乎所有的已知關於 incomplete data 的 MLE 的問題 (或者 Bayesian 的問題), 都可以放在這個架構下討論。使用的市場一大, 哪一個期刊會拒絕它?

February 2, 2007

第十部分：人在江湖

武林弟子藝成之日，拜別師父下山，就進入江湖。以後是否成名立萬，全靠自己的努力。

當你拿到博士的那一天，如果你還預備搞科研，如果你不是去那種不准對外發表的機密單位，那麼你就算是進入了「以發表論文來求生存」的江湖。江湖中有黑道白道。當然，你是名門弟子，一開始算是白道。

白道有白道的規矩。下面所列，不全是白道的規矩。有些是你師父在平常就該教的。如果他忘了，你不妨拿來參考。

不弄虛作假

寫論文的目的是為研究的成果做紀錄，做紀錄的目的是讓後面的人據此可以更向前進。我們所做的這類學問是累積型的¹，牛頓說他的成績，都是因為他能站在巨人的肩上的原故。要看高看遠看深，當然要站得高。

因為是累積型的工作，所以要正確。若你的結果毫無用處，別人根本不會用，就算錯了也不會對科學有影響。但若是你的結果表面甚好，但其中有不能彌補的大漏洞²，別人容易由此可以得到更好的結果，這樣就會有大問題了。

學術論文的基本要求是誠實和嚴謹。準此，論文裡的小錯可以原諒，但不可以故意弄錯，尤其不能在關鍵處故意弄錯³。

引用明確

論文中所引用的文獻，主要以專書和其它學刊期刊的論文為主⁴。至於一般的技術報告，開會的紀錄 (conference proceedings)⁵，專書的一章，網路上所貼的文章，某處的碩士論文等...，這類比較不入流的事物，能避則避⁶——這些都是塗在臉化妝品，你總不希望評審人覺得你在用劣質化妝品罷。

引用論文時，可以不必提你所引用的在文中何處 (有時你還是要提在 section 3.4 或者 equation (4.2) 之類)。但引用專書時，則最好要提章節或頁碼，以方便讀者去查。這一方面是禮貌，一方面則表示「你真的讀過」。

因此，你引用的文獻，應該是你讀過的 (當然不必是全部)。若引用的是冷門資料，有

¹像老子的《道德經》之類，是非累積型的。

²邏輯這回事是有趣的，若是你的錯誤結果被廣泛利用，從這條路走下去，不久就會發現你論文裡有問題。這時，無心的錯或是有心的誤導便一目了然。

³另一個理由是：在關鍵處，這類錯太容易被內行抓住。例如我也在做同一問題，有同一瓶頸，為甚麼你做得出來但我做不出來？競爭者總會用心查一查。

⁴較弱期刊的論文效引用多了，會減少你論文的份量。

⁵這有例外：如 Proceedings of the Berkeley Symposium，雖是開會紀錄，卻是極有聲望的。其它圈內的幾篇有名的技術報告 (如 Efron 在前述 bootstrap 的文章裡所引用的關於 infinitesimal jackknife 的 Bell Labs TM——因為反而讓你顯得內行。) 則不在此限。

⁶因此你就知道為甚麼要少在三流期刊上投稿，因為會較少人引用，而 citation 的數目，暗示你論文重要的程度。

些期刊就會要求你寄一仍複印去 (給評審人參考)。

誰先做出

“Priority” 是學術工作者最在乎的事情之一。因為研究要求的就是求新求變，因此「這件事是誰先做出來的」就不能含糊。例如引用到 jackknife 時，你就需要提到 Quenoulli-Tukey。但為甚麼 Tukey (1958)⁷ 明明在 Quenoulli (1956) 之後，卻要提起？你若是內行，就知道這兩篇文章有基本的不同，而 Tukey 的看法要深刻得多。

準此，你也要保護你的 priority。明明是你做的，被人搶先一步發表，那你的論文 (至少在好期刊上) 就泡湯了。

關於 Priority 的保護，口說無憑。我曾聽到的故事說：有一篇投出的文章被拒，但過了一陣，這個作者被要求評審一篇類似的文章，真是「類似」，因為這篇新文章和他的「舊作」，相差只在他的 α 一律被改為 β 。類似的可怕故事並不是沒有，投稿到有聲望的期刊一般不會遇到這樣事，當你小有名氣之後別人也不見得真敢這樣做。但是，把你的文章壓一壓，同時再做一篇比你的內容更多的文章吃掉你大部分可以做的 extension，卻是可能的。

保護的方法是：要求你的工作單位建立一套有時間、有文號的技術報告系列。這樣，你的工作出來的成績，因為你工作單位的背書，就比較受到保護。例如你大可寄信給期刊的主編、對方的學術領導、或者相關國際學會的理事長等，說你在 1999 年 7 月已做出該結果來，有本校的技術報告為證。這樣的信當然會有影響⁸。如果那人抄得十分明確，他的日子不會好過。因為學術圈子其實不大，這樣的作為會貽笑國際也。

不可硬拗

若是你的文章有錯被別人挑出來，當然第一步是確認是否真是你的錯⁹。若沒錯，當然是頂回去，用詞不必強烈，維持學術上的樸克面孔就好。別忘了還是要感謝別人讀了你的文章。若有不能修改的錯，你也只好承認¹⁰ (並還得在用詞上感謝別人)。

長袖可舞

人際關係永遠是重要的，在學術圈子裡亦然。學術裡會看實力，學問上的和學術資源上的實力。例如假如你手上有一大筆預算支援一些項目，或者你是某不錯期刊的主編，總會多受一點尊重¹¹。沒有辦法——大家都要吃飯。杜甫詩：同學少年皆不賤，五陵裘馬自輕肥。現在大家都差不多，說不定你的學問還好些。但十年後為甚麼有所不同？用心的話，錢愈花愈多，學術資源也是愈用愈多。你若是老啃著那一畝三分地，除非裡面能冒出石油，長遠下去，場面怎會變大？

「在家靠父母，出外靠朋友」。單人隻劍的獨行俠，走江湖可不容易。煮酒放歌，交友論劍，也是白道人物可以做的。

⁷T 的文章，其實只是一個半頁的摘要。

⁸這樣的信，若是由你的單位領導來寫，就有「捍衛本單位人員的學術權益」的意思。

⁹注意到，若是別人有錯在先，你因為引用錯的結果而犯錯，這也算你的錯——誰叫你不察，不夠用功嘛。

¹⁰一般好期刊都允許作者去信更正已發現的錯，這類事當然是愈少愈好。

¹¹但這類事隨著你的權力而來，也隨著你的權力而去。你最好是做莊周的時候用心做莊周，做蝴蝶的時候用心做蝴蝶。

熱心而肯辦事的學者機會會多一點。你若是有這樣的服務性格，不妨多入世一點。有些雜務不妨交給自己的學生做¹²，但這裡有分寸，不要過分。需知你的學生並不傻，這類事英文叫“give and take”，中文是願打願挨。

退一步沒有海闊天空

做研究指的是在第一線做研究。意思是說你總是在做最新的工作，要發明或發現新的東西，而這是要和全世界的一小撮最好的人在比。好期刊不會因為你的行政職位高低來影響是否接受你的論文。即使你曾經在第一線，但學問如逆水行舟，不進則退。你若是有三年的停滯，你就是有三年的停滯。要追回世界水準，也許你還有功力，不需三年。但是你就是要重下功夫才行。

別人也不見得多利害

世上是有一些高手怪傑之類，吾等小民硬是比不過。但不要怕，IQ 的分布大概還是差不多 normal。做研究工作就不能沒有氣性。所謂的「氣不可衰」，前已言及。

多半的人都是在軟木頭上鑽洞。你若是一心搶快，就不要咪咪摸摸的。將學術資源搞起來（如要求本單位訂閱重要期刊），固定去查閱可能的對手有甚麼新的技術報告，儘量去找一些別人沒有的數據……。總之，別閑著。

別一個方法是做較難一點的。對於較難的問題¹³，因為投入/產出比不高，一般精明的搶快型學者是不太肯做的。你若是在偏遠地區，學術資源少，不妨試這條路。當然，你需要注意保護你的 priority。

自己也沒有多利害

意思是，不要選自己做不出來的題目¹⁴。研究工作雖有時候要搶快，但基本上還是要靠穩扎穩打。選題要（1）題目要有一點趣味，（2）至少看起來要有影響，（3）自己做得出來（4）如果你還有餘力，留一點空間給別人做。

February 2, 2007

¹²例如在主辦學術會議時，總需要人辦事。

¹³注意不建議你去做無謂的難題，要選難而有影響的。收尾型的題目往往難而功勞不大，畢竟，把一個房子打掃掉最後的 tough 垃圾再關上門，是無趣的事。

¹⁴你可以問：不去做怎知做不出？Well，用心試一兩個月無妨，但是「知止」也是重要的品行。

附錄 作者自述

個人簡介



退休人士，常建議朋友沒事別寫回憶錄。大頭貼裡的貓芳名“學妹”，很不愛理人的。

在四川巴縣讀幼稚園。以上是我還記得但政府可能查不出來的部分。FB 想要我填的，在大數據時代它一定查得出來，我就不填了。

出生日期

1938 年 1 月 23 日

關於酒店主人

是「酒店關門」後的退休教授。

他又去開一家酒店。這樣，就可以不走，並且只賣愛賣的酒。

關於部落格「趙家酒店」

我想，「部落格」裡面的「格」字，是多餘的。「部落」其實滿好，意思也有一點對。在 blog 裡做版主，與在部落裡做酋長，是不是有一點像？「格」甚麼？

貼了一些現在寫的、以前捨不得丟的東西。太長的文章或者有公式有圖的文章，有些就用 pdf 檔來做了。

比較有用的單元是「關於論文寫作」。也許有用，也許沒有，是留給有緣人看的。