Journal of Propagations in Probability and Statistics (JPPS)
Vol. 1, No. 1, August 2000
pp.127-134

統 計 薪 傳
第一卷 第一期
中華民國 89 年 8 月
127-134 頁

統計大師 William G. Cochran 的事蹟

韓建佩
(Chien-Pai Han)
美國德州大學阿靈頓校區數學系

摘 要

威廉.卡克倫(William G. Cochran)一生都在統計領域中工作，他對統計的貢獻與影響可說是無可計量，而在抽樣調查、實驗設計、observational study 等方面，貢獻尤大，使統計界人士受益無窮。本文簡單地介紹 Cochran 一生的事蹟，同時也將筆者個人與 Cochran 相處的經驗略述一二，以饗讀者。

□本文作者為美國德州大學阿靈頓校區數學系教授暨泛華統計協會現任會長。

統計薪傳 第一卷 第一期 中華民國 89 年 8 月 127-134 頁

1. 前言

William G. Cochran 是統計學界早期的大師，與 R. A. Fisher，J. Neyman，E. S. Pearson 和 Frank Yates 等人齊名。他一生中爲統計界貢獻了許多理論和方法，筆者曾是他的學生，所以對他的事蹟稍有了解。

Cochran 出生於 1909 年 7 月 15 日，生長在 Rutherglen，Scotland。他的許多朋友都稱呼他的小名 Bill。1931 年，他在 Glasgrow 大學拿到碩士學位後就去康橋大學(Cambridge University)唸書，本來要攻讀博士學位，但當時在 Rothamsted Experimental Station 恰有一個空缺，所以他就於 1934 年去 Rothamsted 工作，在那裡和 R. A. Fisher 與 Frank Yates 爲同事。那時 R. A. Fisher 已很有成就，Cochran 跟他學了很多事物與經驗。記得在哈佛大學求學時，有一次筆者和同學們在上 Cochran 的課的時候，大夥兒抱怨 Cochran 給我們太多計算題的習題。但 Cochran 說：「你們現在都有電腦可以用，真好命！以前我和 R. A. Fisher 做事的時候，那個年代沒有電腦，而普通計算機的計算速度又很慢。Fisher 有時要我算一些問題，每次都要花好幾天才能算出來呢！」於是大夥兒就不便再抱怨了。

Rothamsted 是一個農業試驗中心，Cochran 在那裡不但和 Fisher 及 Yates 學了許多統計理論與方法，也增進了不少實際工作經驗，將統計運用在實際問題上。

Cochran 於 1939 年來到美國愛荷華州立學院(Iowa State College，現在已變爲愛荷華州立大學)任數理統計教授。在那裡他建樹很多，發表許多篇論文，同時也和 G. M. Cox 開始合作撰寫「實驗設計」 (Experimental Designs)這本書。後來 G. M. Cox 從 Iowa 搬到 North Carolina State College(現在已成爲 North Carolina State University)，並於 1946 年邀請 Cochran 到 North Carolina 去教書，同時請他擔任該校數理統計研究所的副所長。同年 Cochran 又被選爲 Institution of Mathematical Statistics 的會長。

Cochran 在 North Carolina 待了兩年，然後於 1948 年遷到 Johns Hopkins University 的生物統計系擔任教授。該校的醫學院是美國最好的醫學院之一，常和哈佛大學的醫學院爭第一名。他在那裡有很優越的環境做生物統計的研究。

1957 年哈佛大學要成立統計系，F. Mosteller 邀請 Cochran 到哈佛助陣。Cochran 就搬到 Cambridge，Massachusetts 去了，從此以後他就一直在哈佛大學任教，並於 1976 年退休，成爲哈佛的名譽退休教授(Emeritus Professor)。他退休後住在 Cape Cod，Massachusetts。這裡是一個半島，風景優美，很多很有名的人物如甘迺迪總統的家就在這個半島上，是一個退休的好地方。Cochran 雖然退休但他仍然一直做學術上的研究，一直到 1980 年 3 月 29 日去世爲止。他的名望在統計領域裡真算是無人不曉了。

Cochran 在統計上的成就可以從有些統計方法和定理都用 Cochran 來命名而得到鐵證，其中最有名要算是 Cochran's Theorem、Cochran's test 和 Rao-Hartley-Cochran estimator(以下簡稱 RHC 估計量)。以下各節就將這些方法和他在抽樣調查、實驗設計及其他方面的貢獻做一個簡單的介紹。

2. 卡克倫定理(Cochran's Theorem)

Cochran 發表的第一篇論文是 1934 年的「The distribution of quadratic forms in a normal system with applications to the analysis of covariance」，其內容就是著名的 Cochran's Theorem。這個定理是：「當 n 個變數 $X_1, X_2, ..., X_n$ 都是標準常態分佈，而且都互相獨立時(mutually independent)，如果 $\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{k} Q_i$ ，每個 Q_i 都是非負整數的二次形(quadratic form)， Q_i 的秩(rank)是 n_i ，那麼使 Q_i 的分佈爲卡方(Chi-square)(n_i 自由度)， $i = 1,2,...,k$ ，而且 Q_i 相互獨立的充要條件是 $\sum_{i=1}^{k} n_i = n$ 。

這個定理對統計學的影響非常大，尤其是在方差分析(analysis of variance)和實驗設計方面更是運用廣大。因爲每一個方差分析表中的平方和都可以寫成爲二次形，如果誤差的分佈是常態分布，那麼只要證明 $\sum_{i=1}^{k} n_i = n$ 就可以導致平方差的分佈都是卡方分佈了。

3. Cochran's Test in the Behrens-Fisher Problem

當我們從兩個常態分佈母體中分別抽出兩組樣本，一組從 N(μ_1, σ_1^2)中抽出，另一組從 N(μ_2, σ_2^2)中抽出。假設這兩組樣本是互相獨立的，我們要檢驗虛無假設 $\mu_1 = \mu_2$ 之真僞。如果兩個方差 σ_1^2 和 σ_2^2 不相等，我們就不能用 t 檢驗。那麼要用什麼樣的方法來檢驗呢? 這是個著名的老問題，稱爲 Behrens-Fisher 問題，至今還沒有一個完整的解決方法，尤其是在兩組樣本的樣本數都小的時候，這個問題更是棘手。解決這個問題的方法多半是近似法(Approximation method)，Cochran's test 便是其中之一。此方法最早是在 Cochran and Cox (1950)一書中提出的。Cochran 說大約在一九四幾年的時候，P. V. Sukhatme 在 Sankhya 發表了 Behrens-Fisher test 之 5% 和 1% 顯著水準(significance level)的表格，但是這種表格並不普遍，而且其他百分比之顯著水準又沒有表可以查，所以 G. W. Snedecor 就問 Cochran 是否可以推出一個近似值。Cochran 根據 Sukhatme 的表格提出了可以用 t 分佈表用在 Behrens-Fisher 問題的近似方法。這個方法很簡單，令

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad ,$$

$$s_i^2 = \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i \right) / f_i \quad ,$$

$$f_i = n_i - 1 \ , \ i = 1,2$$

爲兩組樣本之樣本平均數和方差。再令 $W_i = s_i^2 / n_i$ ，而 t_i 爲 t 分布(f_i 自由度)的 $100(1-\%)$ 的百分數點。在抽出樣本後，如果樣本顯示下面的結果

$$\frac{\left| \bar{X}_1 - \bar{X}_2 \right|}{\sqrt{W_1 + W_2}} > \frac{W_1 t_1 + W_2 t_2}{W_1 + W_2} \ ,$$

那麼虛無假設 $H_0 : \mu_1 = \mu_2$ (對 $H_1 : \mu_1 \neq \mu_2$)就被否決。至於 Cochran's test 的 size 和 power，請參考 Lauer and Han (1974)一文。該文中作者們還探討了先檢驗方差相等之虛無假設，再決定是否用 Cochran's test。因爲在實際運用上我們很多時候都不知道 σ_1^2 和 σ_2^2 是否相等。在這種情形下就可以先檢驗虛無假設 $\sigma_1^2 = \sigma_2^2$ 。這種檢驗稱爲初步檢驗(preliminary test)。如果初步檢驗認爲 σ_1^2 和 σ_2^2 相等，那就用 t 檢驗去檢驗 $H_0 : \mu_1 = \mu_2$； 如果初步檢驗認爲 σ_1^2 與 σ_2^2 不相等，那就用 Cochran's test。當然初步檢驗對 size 和 power 都有影響，這就是我們對初步檢驗要特別注意的地方。有關初步檢驗之研究論文極夥，有兩份重要目錄如後: Bancroft and Han (1977)與 Han, Rao, and Ravichandran (1988)，在此順便提出，供讀者參考。

說到檢驗方差，Cochran (1941)還研究檢驗幾個方差同時相等的問題。如果有 k 個方差而要檢驗虛無假設 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ 時，Cochran's statistic 是

$$C_{\max} = \frac{\max(s_1^2, s_2^2, ..., s_k^2)}{s_1^2 + s_2^2 + \cdots + s_k^2} \quad 。$$

這個檢驗方法在一個 σ_i^2 很大而其他(k-1)個 σ_i^2 都相等的情況下最爲有效。如果這個條件不適合，一般都用 Bartletts test 去做檢驗。

4. RHC 估計量

Rao,Hartley, and Cochran (1962)一文中提出的 RHC 估計量是用在不等機率不置回抽樣(unequal probability sampling without replacement)的一種方法。雖然不等機率抽樣方面有許多方法，如 Horvitz and Thompson (1952)、Yates and grundy (1953)等。但這些方法要計算樣本被選的機率(inclusion probability)，而這些計算公式都非常複雜，所以在實用上有困難。Rao、Hartley 和 Cochran 就提出一個簡單的方法，這個方法可以算出母體總額的不偏估計值(unbiased estimate of poulation total)，同時這個估計值的方差及該方差的不偏估計值都可以算出來，這樣在實際運用上就簡單多了。

RHC 估計量可以簡單敘述如下:在一個母體中有 N 個個體，我們從中抽出 n 個個體，每一個個體被抽中的機率是 P_i ， $i = 1,2,..., N$ ， $\sum P_i = 1$ 。這個方法先將 N 個個體用隨機方法分爲 n 個小組，每一個小組有 N_j 個體， $\sum_{j=1}^{n} N_j = N$ 。然後從每一個小組中依據 P_i 的值抽一個個體，這樣有 n 個小組，所以樣本中就有 n 個個體了。

如果在第 j 個小組中將被抽中之的個體的觀測值寫爲 y_j ，其機率爲 P_j ，同時令 g_j 爲第 j 小組中所有 P_i 的總和，那麼母體總額 $Y = \sum_{i=1}^{N} y_i$ 的 RHC 估計量就是

$$\hat{Y}_{RHC} = \sum_{j=1}^{n} \frac{y_j}{P_j / g_j} \quad ,$$

而 $\hat{Y}_{RHC}$ 的方差爲

$$V(\hat{Y}_{RHC}) = \frac{n\left(\sum_{j=1}^{n} N_j^2 - N\right)}{N(N-1)}\left(\sum_{i=1}^{N} \frac{y_i^2}{nP_i} - \frac{Y^2}{n}\right) \quad 。$$

這個方差的不偏估計量爲

$$\hat{V}(\hat{Y}_{RHC}) = \frac{\left(\sum_{j=1}^{n} N_j^2 - N\right)}{\left(N^2 - \sum_{j=1}^{n} N_j^2\right)}\left(\sum_{j=1}^{n} g_i\left(\frac{y_i}{P_i} - \hat{Y}\right)^2\right) \quad 。$$

5. 抽樣調查

除了 RHC 估計量之外，Cochran 在抽樣調查方面還有許多貢獻。他的教科書 Cochran (1977)「Sampling Technique」從 1953 年發行到現在已修定至第 3 版 (1953，1963 及 1977)，一直是學抽樣調查必讀的聖經。當然 Cochran 也用這本書教他的學生。他在上第一堂課時一定先談爲什麼要抽樣，主要的考量因素是(1)經費:因爲抽樣只需要收集一些樣本的資料，比普查要便宜得多，而且一般在做調查研究時總會有一定的預算，因此就可以依預算大小用最好的方法去抽樣與求估計值。(2)時間:當然抽樣比普查省時，尤其民意調查都有時間限制，過時的資料就不會有多少用途。如預測選舉的結果就是很好的例子。(3)人力資源:在許多調查方面要用經過嚴格訓練的人員。這些人的人數有限，所以只能做抽樣調查，無法用普查。(4)準確度:一般普查的工作量都會很重，那麼誤差就多了，如果用抽樣調查，資料較少，處理也就完整些，準確度就加大了。在討論過爲什麼要抽樣後，Cochran 就會指出抽樣一定會產生誤差，這些誤差可以分爲以下幾類:

1. 不完整誤差(Incompleteness)
   A. 誤差產生於無記錄(Noncoverage)
   B. 誤差可從無回答(Nonresponse)而產生
2. 衡量誤差(measurement error)
3. 抽樣誤差: 因爲抽樣只是母體的一部分，所以抽樣誤差是不可避免的。
4. 時間誤差: 母體可因時間不同而有改變，所以樣本就會有誤差。

以上這些誤差都要用有效的方法使其愈小愈好。Cochran 的 Sampling Technique 一書針對以上各種誤差都有說明。這本書還介紹許多抽樣方法，如簡單隨機抽樣法、分層隨機抽樣法、等距(系統)抽樣法、集群抽樣法等等。Cochran 在教課時會用淺近的例子把統計方法傳授給學生，這點在他的書中也很明顯，因爲書中有許多實際運用的例子，可做參證。

6. 實驗設計

Cochran 到 Rothamsted 在 F. Yates 那裡工作，因而得到許多實際工作經驗，這對他後來和 G. Cox 合寫「實驗設計」(Cochran and Cox 1950)這本書很有幫助。由於 Rothamsted 是一個農業試驗所，而後來 Cochran 到 Iowa，那裡也是一個農業州，所以他的實驗設計一書中農業方面的例子很多。

註解 [Frank1]:

這本書的第一版是在 1950 年發行，發行後就一直盛銷。第二次修正版是在 1957 年發行，書中包括許多不同的設計方法，如完全隨機設計(completely randomized design)、隨機區段設計(randomized block design)、拉丁方格設計(Latin square design)、因子實驗(factorial experiments)、不完全區段設計(incomplete block design)等等。雖然這本書是參證書沒有習題，但是許多老師，包括 Cochran 本人，都用這本書做教科書。

Cochran 不但寫實驗設計理論上的文章，也寫一些實用的文章，例如在 Cox and Cochran (1946)這篇論文中，就談到如何在溫室中做擺設，才可以增進實驗設計的效果。

7. 從實際問題上著手

Cochran 發表的論文有許多都和實際應用有關，他可以將實際問題轉變爲統計問題，然後尋求適當的解答。 Cochran 早期在 Rothamsted 工作就有這方面的訓練和經歷，後來他去 Iowa State University 教書，那裡也有良好的統計實務應用環境。Iowa 是一個農產品中心，尤其玉蜀黍的產量爲全美國之冠。所以在統計的應用方面對農產品的增產和估計等問題往往都是一慣的研究對象，這也是爲什麼在 Cochran 寫的的書中有許多農業方面的例子。在 Iowa 統計系教書的老師有一句笑話，那就是如果上課時不用玉蜀黍做例子，就不能算是 Iowa State 的老師。還有一句笑話是，如果你給一個專題演講(seminar)而且你能一開始就將 $Y = x\beta + \varepsilon$ 這個式子寫在黑板上，那就一定會受到聽眾的歡迎。大概這些都可能是 Cochran 到 Iowa state 之後留下來的傳統吧!

Cochran 到 Iowa state 的時候，G. W. Snedecor 是那裡的統計圈領導人物。Snedecor 寫了一本書 Statistical Methods，是統計學中最暢銷書之一。Snedecor 去世後，Cochran 將這本書修改，加入許多抽樣調查及實驗設計的資料，這本書就成爲 Snedecor and Cochran (1967)名作，仍爲暢銷書，而且被翻譯成許多國文字，而裨益於世界各國學統計的人。

Cochran 的實際研究工作不止於農業方面，在醫學方面他也有許多貢獻。Cochran 在 Johns Hopkins 教書時，他接觸的問題多半是醫學方面的，所以他在生物統計方面寫了許多文章，最有名的可能要算 Kinsey 報導的統計問題(Cochran 1953 以及 Cochran, Mosteller, and Tukey 1954)。因爲這個報導的內容與性生活方面的問題有關，所以很令人觸目。當然這個報導是以討論統計方面的問題爲主。

至於在其他生物統計方面，1964 年 Cochran 和其他幾個研究人員，寫了一個報告給美國衛生署長(Surgeon General) ，報告的內容是抽煙與許多病症的關係。這個研究問題是屬於 observational study(觀察研究)，參入研究的對象有好幾百萬人，研究了好幾年，因此有些人在中途就退了出來了，這樣就會造成研究結果有偏差。Cochran 就想辦法將這種偏差矯正過來，使估計更爲準確。

Cochran 在 observational study 方面做了很多研究，而這類研究中常常會有混同因子(compounding factor)在內，例如抽煙和癌症會有關聯，可是抽煙不是導致生癌症的唯一原因。一個人患癌症會有許多不同的因素，如遺傳、年齡、生活環境等，在做統計研究時對這些因素都要顧及到。

Cochran 提出兩個主要方法來控制混同因子和偏差。一個方法是將母體分爲幾個部份，如根據年齡分組，然後在每一部份個別做研討。另一個方法就是用協方差矯正法(covariance adjustment)，一般都可以用迴歸法去做矯正。

8. 結論

W. G. Cochran 在統計學上貢獻良多，以上只是簡略地介紹他一生的成就。其實他除了在學術上有卓越的貢獻外，他的爲人處世也可做爲後人的榜樣。

Cochran 爲人謙和，尤其很愛護他的學生。在學業上儘量幫助學生學習，指導他們研究方向，提供他們研究題目並幫助解決研究上的困難。筆者身爲他的學生，親身受過他的恩澤，因爲我的博士論文題目就是 Cochran 提供的。他那時去 Sabatical leave，我的論文就在 A. P. Dempster 教授指導下完成。後來 Cochran 休假回來，我就將論文交給他看，也得到他的批准。

Cochran 帶了許多學生做論文。他在 Iowa State University 的時候指導過 T. A. Bancroft 的博士論文。這篇論文是討論初步檢驗的問題，後來在 Annals of Mathematical statistics(Bancroft 1944)上發表，成爲初步檢驗問題的第一篇始祖文章，所以在初步檢驗的問題上 Cochran 也有一份貢獻。

我覺得做 Cochran 的學生是很幸運的，因爲他上課時雖很嚴肅，可是下課後則較隨和，也會在走廊上和學生聊天。有一次我在走廊上和他聊天，我知道那時他正在研究抽煙和其他病症的關係，於是就問他爲什麼還在抽煙而不戒煙，他說這是他個人的選擇，因爲他知道抽煙不是導致癌症的原因，只是兩者之間有關係而已。

Cochran 一生寫了六本書，同時還發表了 116 篇文章，這些文章已收集成一本書，書名爲 Contributions to Statistics。Cochran 的朋友、同事、以及他的許多學生也合寫了一本紀念他的書，這本書由他的兩個門徒 P. S. R. S. Rao and J. Sedransk (1984)擔任編輯邀請大家合寫，書名爲 W. G. Cochran’s Impact on Statistics。這也就可以聊表大家對 Cochran 的敬意了。

參考文獻

Bancroft, T. A. (1944). On biases in estimation due to use of preliminary tests of significance. *Annals of Mathematical Statistics*, **15**, 190-204.

Bancroft, T. A. and Han, C. P. (1977). Inference based on conditional specification: a note and a bibliography. *International Statistical Review*, **45**, 117-127.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal System. *Proc. Camb. Phil. Soc.*, **30**, 178-191.

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Eugenics*, **11**, 47-52.

Cochran, W. G. (1953). Statistical problems of the Kinsey Report. *Journal of the American Statistical Association*, **48**, 673-716.

Cochran, W. G. (1964). Approximate significance levels of the Behrens- Fisher test. *Biometrics*, **20**, 191-195.

Cochran, W. G. (1977). Sampling Techniques,3rd ed.,. John Wiley, New York.

Cochran, W. G. and Cox, G. M.(1957). Experimental Designs,2nd ed.. John Wiley, New York.

Cochran, W. G., Mosteller, F., and Tukey, J. W. (1954). Principles of sampling. *Journal of the American Statistical Association*, **49**, 13-35.

Cox,G. M. and Cochran, W. G. (1946). Design of greenhouse experiments for statistical analysis. *Soil Science*, **62**, 87-97.

Han, C. P., Rao, C. V., and Ravichandran, J. (1988). Inference based on conditional specification: a second bibliography. *Communications in Statistics-Theory and Methods*, **17**, 1945-1964.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

Lauer, G. N. and Han, C. P. (1974). Power of Cochran's test in the Behrens-Fisher problem. *Technometrics*, **16**, 545-549.

Rao, J. N. K., Hartley, H. O., and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B*, **24**, 482-491.

Rao, P. S. R. S. and Sedransk, J. (1984). W. G. Cochran's Impact on Statistics. John Wiley, New York.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B*, **15**, 253-261.