種類知多少?
.... 敬獻給為台灣環保努力的朋友們

趙蓮菊

1. 前言:

一般所謂的「生物統計」其研究的主題可爲人類或動、植物。例如統計應用在醫學、流行病學及遺傳學時, 其主要研究對象是人類; 而本文所討論者是統計應用於生物學或生態學, 而基本的研究對象是動物和植物。在後者的應用中, 一個極爲古老和有趣的問題即是種類數的估計。

首先舉一個實際的例子來說明。當我們觀察清華園中鳥的種類時, 有些鳥種非常普遍, 極爲容易看到, 例如麻雀、白頭翁、綠繡眼、小白鷺和金背鳩等; 但有些種類卻是驚鴻一瞥, 難得一見的, 例如領角鴞、赤腹鷹及黑枕藍鶲。(參考"水木飛羽話清華"一書)。截至1991年4月爲止的記錄中, 共有28科69種。但是否這69種就是清華園中所有的鳥種類呢? 顯然不能確定, 因爲各種鳥種都是經年累月, 陸陸續續被鳥友觀察發現到的, 那麼到底在清華園內還有多少種尚未被觀察到呢? 有沒有辦法從已經看到的種類記錄去推算尚未出現的種類數呢? 這就是本文的主題。

種類數的估計往往因調查地區的地形不同而有極大的差異。例如就在離清華校園不遠的客雅溪口, 早期是台灣西岸的重要濕地, 近年來人爲的污染, 水泥堤的興建已造成嚴重的生態破壞, 但仍是許多水鳥的棲息地。新竹野鳥學會的記錄中在此地區已有155種鳥種, 光是鷸科就有27種之多, 其中數量最多的是濱鷸, 1994年4月至次年4月就曾看到約5013次總數, 然而濱鷸幾乎不曾在清華園中出現過。我們以新竹野鳥學會1994年4月至次年4月在客雅溪口的65次的調查數據爲例來說明一般資料的形式。(新竹野鳥學會資料提供)。在所觀察到的155種中有96種其出現的次數超過10次, 這些鳥種基本上是此地普遍且容易觀到的鳥種; 而不超過10次出現的鳥種有59種, 這些是比較稀有的鳥種。我們現在有興趣的問題是估算那些不曾被觀察到的鳥種數。當然這些未曾被觀察到的鳥種一定是屬於較稀有的種, 因此那些出現許多次的普遍種幾乎不能提供任何訊息, 所以我們主要利用已出現的較稀少的種數來估計不曾觀察到的鳥種數。基本上就是利用出現不超過10次的那些資料來作此分析。此組資料如下列

出現次數 (i)	種類數 (f_i)
1次	25種
2次	6種
3次	10種
4次	3種
5次	4種
6次	4種
7次	1種
8次	1種
9次	4種
10次	1種
不超過10次的共有59種	
超過10次以上者共有96種	

現在的問題是在調查中不出現 (即出現0次者) 的種類數有多少? 這裡選擇不超過10次的資料基本上完全由經驗得來。對於此組數據, 我們將在第4節中分析。

2. 相關的其它應用

前述所提到的「種類數估計」和「總數估計」有很密切的關係。在種類數估計中由於可從外觀分辨出不同種類, 因此在鳥類調查時我們以眼睛或望眼鏡看到鳥後, 即可記錄下鳥種及隻數, 這些鳥可以是重複看到的。但現在如果我們的目的是要估計某一特定種類的「總數」時, 例如我們想知道黑面琵鷺共有幾隻時, 只須將每一隻都視爲一個特別的「種類」。但由於所有的黑面琵鷺外觀均相同, 不易分辨出個別約一隻, 因此必須在每一隻觀察到的琵鷺上作「記號」, 例如套上脚環以資辨別, 這就是「繫放」或「重複捕取」的原理。我們可利用重複捕取到的比例多寡來估計未看到的隻數, 一如我們利用種類出現的次數來估計未看到的種類數一樣。因此, 這二個問題在統計分析的理論來看是完全類似的。

然而所謂「種類」的定義並不一定侷限於在生物學上爲人所熟知的生物分類。在一般的實際應用中,「種類」可以有很廣泛的定義。以下舉一些作者所知的應用與其對應之「種類」的意義:

(1) 估計人口普查之短缺人數: 一般所謂之人口普查應該調查到每一個母體中的人, 但事實上大規模的全國人口普查幾乎都漏掉了一些人。國內目前尚無因普查短缺而引起紛爭, 但在美國, 以紐約市爲例, 即曾上法庭控告美國普查局少算了紐約市的人數以致可能影響了紐約市所分配的預算及國會席位。(見 Fienberg, 1992及其中所列相關之一系列之文章。) 但要如何去估計短缺的人數呢? 顯然只由普查的數據是無法知道的。由此在美國每十年一次的人口普查完後, 又作一次「事後調查」, 利用重複的訊息來估計缺少的人數。在此應用中, 所謂之「種類」即爲母體中不同之個人。

(2) 在流行病學上估計罹患某一種病的總人數: 在流行病學中, 罹患人數是用來得到罹病比率的最基本資料。但一般流行病學中的罹患人數即爲「調查到的罹患人數」。事實上可能有一些罹病者沒有被調查到, 此地的「種類」即爲罹病的個人患者。例如在 Fienberg (1972) 論文中欲估計某地區患唐氏症之嬰兒數, 利用5種記錄來作爲「重複捕取」的資料。這5種記錄分別是醫院婦產科記錄、

住院記錄、公共衛生部門記錄、精神健康部門記錄以及特殊殘障學校的記錄。這五種記錄中共有537個嬰兒, 有許多是有重複登記的。如果利用重複出現的次數去估計在此5個記錄中的均無出現的嬰兒數, 即可得到罹患此病的總人數估計。

(3) 估計軟體中錯誤的數目: 軟體的可靠度可經由軟體中錯誤的數目來測度。通常一個軟體中有些錯誤在固定時間內被偵測到許多次, 而有些錯誤可能只被偵測到數次或根本無法偵測到。因此若將「種類」視爲軟體中之不同錯誤, 我們的問題即爲利用偵測到之錯誤之出現次數來估計沒有被偵測到的錯誤有多少。

(4) 估計某種稀珍郵票現存之數量: 現假設「種類」定義爲某種郵票中的每一單張, 利用每一單張在郵票拍賣場重複出現的次數即可估算同型郵票不曾出現在拍賣場的大約有多少張, 進一步得知現存的總數量。

(5) 估計出土錢幣的鑄模數: 這是在經濟學上的一個應用。當一批古錢幣出土時, 經濟學家希望經由錢幣在市面上的流通量來推測當時代社會經濟的繁榮情況。由於鑄造錢幣的鑄模能鑄造錢幣的數目大致知道, 因此只需估計當時有多少的鑄模即可推算流通的錢量。此地「種類」即爲不同式樣設計的鑄模。例如古代錢的正面通常都是比較複雜的設計, 大部份爲君王的肖像, 而錢的反面則爲較簡單的花樣設計。我們由正、反二面不同鑄模出現的次數可以分別估計正、反二面未出現的鑄模多寡, 進一步估計錢幣之種數。

(6) 估計某一作者所知的字彙數: 例如唐代平民詩人白居易用字平易近人, 著作極多, 但他一生所寫的詩中用了多少個不同的字? 當然我們現在只要鍵入所有他的作品, 即可利用電腦算出。但從統計的眼光來看, 我們只需抽一部份他的作品, 即可由此部份不同字出現的頻率數目來推算沒有抽到的部份尚有多少不同的字。因此在此應用中, 「種類」係指全部作品中之不同字數。國外一個有名的應用即是估計莎士比亞所知的字彙數。莎士比亞一生的全部作品共有884647個字, 其中有許多重複的字, 而共有31534個不同字彙。假如我們假想莎士比亞的作品爲其腦中所知字彙的表現在寫作上的一個樣本, 則還有多少字彙是他所知而沒有表現在作品上的? 此時「種類」數爲莎士比亞腦中所知的不同字彙數。我們將此組數據中出現次數不超過10次的字彙數列於下:

出現次數 (i)	字彙數 (f_i)
1	14376
2	4343
3	2292
4	1463
5	1043
6	837
7	638
8	519
9	430
10	364
不超過10次總字彙數: 26305	
超過10次之字彙數: 5229	

在第四節中我們將以此數據爲例來分析。完整的資料可參考 Efron & Thisted(1976)。

3. 如何估計?

其實估計的方法很多種, 在1993年出版的一篇回顧此主題的論文(Bunge & Fitepatrick) 中敍述有約550篇的論文。無論是有母數假設的估計法、無母數估計法, 或者是貝氏作法, 頻率論作法都應有盡有、「種類」繁多。本文僅介紹作者於1992年所提出的一種方法 (Chao & Lee, 1992)。此種方法基本的原理是: 若要直接估計種類數是很困難的, 但一般「樣本涵蓋」卻可估計的很精準, 因此利用樣本涵蓋率的估計值來間接推算種類數。所謂「樣本涵蓋」C(sample coverage) 之定義如下:

$$C = \sum_{i=1}^{N} p_i \times I[\text{第}i\text{個種類在樣本中至少出現1次}],$$

此地 N 爲未知的種類數, p_i 爲第 i 個種類在全部母體中所佔之比例, 也就是當任取一個觀察值它會屬於第 i 個種類之機會, 而 I(A) 爲事件 A 之指示函數, 即若 A 發生 (不發生) 則 $I(A) = 1(0)$。因此「樣本涵蓋」即爲觀察到之種類機率和。樣本涵蓋隨樣本改變而隨之改變。嚴格說來, 它應該是參數 $\{p_1, p_2, \cdots, p_N\}$ 之隨機變數。它的一個最常使用的「估計量」爲

$$\hat{C} = 1 - f_1/n,$$

此地n爲觀察樣本總數, f_i爲出現i次之種類數。此估計量最早的發現可溯回至計算機始祖 Turing(參考 Good,1953), 此估計量估的極爲精準, 而且僅用到出現一次的種類數目。

假設所有的種類觀察到的機率都相同, 即$p_1 = p_2 = \cdots = p_N = 1/N$, 那麼很顯然$C = D/N$, D爲樣本中出現之不同種類數。所以我們馬上得到若種類機率沒有大小之分, 則種類數的估計量爲

$$\hat{N}_0 = D/\hat{C} = D/(1 - f_1/n)。$$

假若有些種類較易觀察到(如普遍的鳥種) 有些則很難觀察到(如稀有的鳥種), 表示種類之間有大小之變化, 其變化的程度可以「變化係數」γ來測度, 其中

$$\gamma^2 = N\sum_{i=1}^{N}\left(p_i - \frac{1}{N}\right)^2 = N\sum_{i=1}^{N} p_i^2 - 1。$$

當 $\gamma^2 = 0$ 即代表 p_i 之間沒有變化, 也就是回到上面 $p_1 = p_2 = \cdots = p_N = 1/N$ 的情形。若 $\gamma^2 > 0$ 則其值愈大, 表示 p_i 之間變化的程度也愈大。此種情況下種類數的估計量爲 (詳細導證細節, 請參考 Chao & Lee, 1992)

$$\hat{N} = \frac{D}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}}\,\hat{\gamma}^2,$$

其中 $\hat{\gamma}^2$ 爲 γ^2 之一估計量, 具有下列形式

$$\hat{\gamma}^2 = \max\left\{\hat{N}_0 \sum_{i=1}^{n} i(i-1) f_i/[n(n-1)] - 1,\ 0\right\}。$$

事實上, 變化係數的估計是本方法中最困難的一個環節。若種類出現次數較少時, $\hat{\gamma}^2$ 多半會低估, 但假使有部份爲出現很多次的普遍種時, 又產生高估的情形。對於高估的一個簡單的調整已在第一節描述, 即爲僅考慮出現次數不超過10次以上的種類來估計未出現的種類數, 因爲出現多次的普遍種很難含有

未出現種類的訊息。而對於一般低估的情形, 我們採用一個消去偏差的估計量如下:

$$\tilde{\gamma}^2 = \max \{\hat{\gamma}[1+n(1-\hat{C}) \sum i(i-1)f_i/[n(n-1)\hat{C}], 0\}$$

而得到下列的一個種類數估計量:

$$\tilde{N} = \frac{D}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}}\tilde{\gamma}^2。$$

一般在分析數據時, 首先檢查 $\hat{\gamma}^2$ 之值, 若 $\hat{\gamma}^2 = 0$, 表示機率相同的假設可以接受, 於是 $\hat{N}_0$ 便是一個合理的統計量。若 $\hat{\gamma}^2 > 0$, 表示機率之間有大小差異, 則建議使用 $\hat{N}$ 或 $\tilde{N}$。至於如何選擇, 在過去的一些模擬實驗中 (Chao & Lee,1992; Chao, Ma & Yang, 1993) 各有優劣表現, 目前還沒有很明確的方法用來判斷究竟應該使用 $\hat{N}$ 或 $\tilde{N}$。至於估計量的標準差已可由一般的極限方法得到近似估計值。作者並有一套軟體程式計算所有本文所提到的估計量及其近似標準差, 可來函索取。

4. 例題解說

例一: 客雅溪口種類數分析

此數據列於第一節中, 我們針對出現1∼10次來分析, 共有59種不同種類, 即$D = 59$。樣本數 $n = \sum_{i=1}^{10} if_i = 184$, 故樣本涵蓋的估計值 $\hat{C} = 1 - f_1/n = 1 - 25/184 = 86.4\%$。假若種類機率可假設爲相等, 那麼種類數約爲 $\hat{N}_0 = D/C = 59/86.4\% = 68$。問題是種類機率相等是否爲一個合理的假設呢? 我們來計算一下變化係數平方的估計值: $\hat{r}^2 = 0.5897$ 而另一估計值爲 $\tilde{r}^2 = 0.987$。因此很顯然機率相等的假設是不合適的。換言之, $\hat{N}_0$ 將會低估。低估的大小約爲 $\hat{N}$ 或 $\tilde{N}$ 中具有變化係數的該項, 亦即 $f_1\hat{r}^2/\hat{C} = 17$ 或 $f_1\tilde{r}^2/\hat{C} = 29$。所以比較合理的結論是估計量應爲 $\hat{N} = 85 \pm 11$ (11爲標準差) 或 $\tilde{N} = 97 \pm 18$。由此可得在當地從來沒有觀察到的鳥種約有$85 - 59 = 26$至$97 - 59 = 38$種。而客雅溪口的種類數共有$85 + 96 = 181$ (96種爲出現10次以上較普遍種), 或$97 + 96 = 193$ 種。換言之, 如果我們再經年累月的努力觀鳥, 期待會有26-38種新鳥種會被發現。

例二: 莎士比亞字彙數分析

根據第二節中莎士比亞用字的數據, 在出現1∼10次中共有62155個字, 而其中有26305個不同字, $D = 26305$, $n = 62155$。樣本涵蓋估計值爲$\hat{C} = 1 - 14376/62155 = 76.87\%$, 其意義爲莎士比亞表現於寫作字彙的機率和大約只有2/3左右。由於在文字學中, 有許多字是經常使用的, 例如介系詞的of, 或冠詞之 the,a, 而有些字顯然是較艱澀、冷僻, 因此直覺上字彙出現機會相等就是不甚合理, 也就是說$\hat{N}_0 = 26305/76.87\% = 34220$必定是嚴重低估的。變化係數平方之估計值爲$\hat{\gamma}^2 = 0.8207$ 及 $\tilde{\gamma}^2 = 1.6373$, 也直接說明了必須考慮字彙機率不等之估計值$\hat{N}$或$\tilde{N}$。首先以$\hat{N}$來說明, $\hat{N} = 49568 \pm 428$。因此莎士比亞所知但未出現在其作品中的字彙約爲$49568 - 26305 = 23263$個字, 而其總共所知的字彙約爲$49568 + 5229 =$

54797。其次若以$\tilde{N} = 64840 \pm 865$來看, 未出現在作品中之字彙約有38535個, 而其總共所知之字彙約有7萬個, 此與 Efron 和 Thisted 結果很類似。

5. 後語

我在研究種類數估計的多年期間, 隨著我的鳥會朋友們各地賞鳥, 總是驚慌的發現台灣各地生態環境的日益惡化, 人爲的破壞使得處處成爲「傷心地」。謹以此文敬獻給所有爲台灣環保努力的朋友們。

參考資料

1. 清華大學 (1992) 水木飛羽話清華, 李雄略及新竹野鳥學會黃麟鵬主編。
2. Bunge, J. & Fitzpatrick (1993). Estimating the Number of Species: a Review. J. Amer. Statist. Asso. 88, 364-373.
3. Chao, A. & Lee, S-M (1992), Estimating the Number of Classes via Sample Coverage. J. Amer. Statist. Asso. 87, 210-217.
4. Chao, A., Ma, M-C & Yang M. C. K. (1993), Stopping Rules and Estimation for Recapture Debugging with Unequal Failure Rates. Biometrika 80, 193-201.
5. Efron, B. & Thisted, R (1976). Estimating the Number of Unseen Species: How Many Words Did Shakespeare know? Biometrika 63, 435-447.
6. Fienberg, S.E. (1972). The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables. Biometrika 59, 591-603.
7. Fienberg, S.E. (1992). An Adjusted Census in 1990? The Trial. Chance 5, 28-38.
8. Good, I.J (1953). The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika 40, 237-264.

—本文作者任教於清華大學統計學研究所—