數學傳播　31卷2期, pp. 3-20

統計裡的估計

黃文璋

一. 前言

繼購物贈送 Hello Kitty 磁鐵活動後,7-ELEVEN 在民國94年的第四季, 推出贈送迪士尼公仔的活動。共有42個一般型的公仔, 外加10個白金色的公仔。我將原先四斗櫃上的雜物清除掉, 然後將那些已擁有的公仔依序排好, 重複的則排在同一行。7-ELEVEN 差不多是我每天必去的店, 只要購物每滿88元, 就會拿到一個公仔。要打開盒子才知到底拿到那一個, 頗有樂趣。假設這些公仔是隨機發放, 則要收集齊全, 所需總個數之期望值是可以求出的。這是所謂古典票券收集問題(classical coupon collector's problem), 這裡面有一些有意思的機率統計問題。隔一段時間後, 終於集全了, 我仍繼續將它們整齊地排好, 櫃子上像有一群兵馬俑一般, 每行長短不一。

有一天唸數學系大四的女兒來我研究室, 對那一片公仔頗好奇。她從小對收集這類小東西很感興趣, 長大後不熱中了, 換成他老爸。我突然問她“總共有幾個, 給妳5秒”。“65個”, 她脫口而出。“不太有概念喔! 妳看, 這42種每一種多的有10個, 少的才2個, 平均幾個? 大約3到4個。如果假設4個, 則共168個, 而那些白金的都很少, 一眼望去全部才12個, 如此共180個。如果假設那42種, 平均每種有3個, 則共126個, 加上12個白金的, 共138個。所以應共有138至180個。” 我對女兒這樣說。我們數了一遍, 結果是143個。

請幾個朋友到餐廳用餐, 分別點了不同的套餐, 價格不一。甜點用完後, 侍者拿來帳單。總共17,270元, 其中還包含小菜、開瓶費及1成服務費等。有沒有多列或算錯呢? 請客嘛, 總不好意思一筆一筆對。四捨五入算至百位, 而有些套餐是一樣的價格, 大致乘一乘, 估算出來的總和, 與帳單上的和相比, 差異在500元左右, 可以接受。

自兩千多年前起, 埃及人就已在估計地球、月亮及太陽的大小。處在此一隨機世界, 我們可說經常在估算、估計。拿起一個蘋果, 估計它的重量。對初認識的朋友, 估計他的年齡, 估計其家庭狀況。要過馬路, 黃燈已亮, 估計是否過得去。有些情況估計準不準不太重要, 可以隨便

估估。有時就會稍留意些。例如, 收到大學指定科目考試成績單, 要做落點預測。各校系選擇方案有別, 採計科目也不同, 有些科目還加重計分, 想估計能考上那一校系, 就頗具挑戰性。其他如估計歌手周華健全年總收入, 估計去美國留學的費用等。我國內政部還訂有“地價調查估計規則”。科學家及政府, 也常在做種種估計。如

1. 估計宇宙年齡、地球重量、全球人口數。
2. 估計大陸石油蘊藏量。
3. 估計海棠颱風對台灣造成的農林漁牧業損失。
4. 估計一銅板出現正面的機率。
5. 估計10歲孩童身高的分佈。

這類估計顯然都不太容易。對於隨機現象, 既然是估計, 便可能有誤差。就如謝霆鋒主唱的那首“估計錯誤”, 其中的一句歌詞: “計算自己情感, 偏偏估計錯誤”。一般人以爲對自己總該較了解, 但似乎仍易估計錯誤。更何況對那些未知的事物, 或多變的隨機現象, 所做的估計, 誤差當然是難免。像是有科學家估計宇宙的年齡介於100億至160億年間。這類估計, 不論依據爲何, 究竟有多精確, 是很難“估計”的, 誰也不知正確答案。所以每隔一段時間, 總會有人提出新的估計值。即使對地球上現存的事物, 其估計也常無定論。例如, “科學人”雜誌的“發表新鮮事”專欄, 2003年8月1日, 刊登鄭靜琪所撰的“鯨魚數量已經安全無虞?”一文。其中寫道, 新的研究指出北大西洋座頭鯨數量的歷史高峰應有24萬, 據此推算, 全球座頭鯨的數量曾多達150萬, 遠遠高於國際捕鯨委員會 (International Whaling Commission) 所估計的10萬頭。另外, 有位哈佛大學 (Harvard University) 的教授, 依據紐約市51家醫院, 所提供的3萬名病人資料, 發現其中約有 4% 的病人因醫療不當而受害。他因此推算美國一年死於醫療疏失的病人有10萬人。你可能會好奇: 紐約市的資料能代表全美國嗎? 事實上常有人是如此以局部放大到全體。如 BBC 中文網, 2006年6月19日有一則“中國野生熊貓比估計多兩倍”之報導:

> 科學家根據新技術推斷, 中國目前的野生大熊貓數量可能達到三千隻, 比過去估計的多了兩倍。一直以來科學家都是用傳統方法估算野生大熊貓的數量, 估計總共有一千多隻。但是鑒於大熊貓獨特的生活習性和生存環境, 人們很難做出準確的推算。中國和英國的研究人員現在分析大熊貓糞便 DNA, 推算熊貓的數量。他們估計四川王朗自然保護區目前應該有66隻野生大熊貓, 比1998年推斷的多了至少兩倍。...
>
> 科學家根據王朗自然保護區的研究結果推斷, 目前中國全國大約有2500到3000隻野生大熊貓。...

我們不淸楚如何經由分析糞便 DNA 以估計熊貓數量。而由一個地區的估計, 是否適合放大到全國, 也不太肯定。因此日後若有人提出新的估計, 將不會令人太驚訝。

估計的方法很多。民國94年12月, 打過大聯盟的棒球好手陳金鋒, 回台灣加入 La new 熊隊。熊隊當然士氣大增, 總教練洪一中說“我不能告訴你明年 La new 可以拿第幾名, 但每年都預估要拿第一”。賣西瓜者用指頭敲一敲, 便可估計出西瓜甜不甜。他的估計是基於什麼? 聲音大小嗎? 考完試, 也沒對答案, 有人便估計自己可考上台大醫學系, 所憑藉的可能是過去在學校裡考試的表現, 加上對這次考試的感覺。估計可以很科學, 也可以只基於主觀或直覺的判斷。但主觀或直覺, 如果再加上一些客觀的數據, 常可讓估計更精準些。洪一中的估計, 可能是衡量各隊實力後客觀的判斷, 也可能是爲了展現信心而高估, 如此則並不需要提供什麼佐證。附帶一提, 有了陳金鋒的加盟, La new 隊果然在民國95年6月28日, 獲得隊史上首座季冠軍。民國92年上半球季, 他們的戰績是慘不忍睹的9勝37敗4和, 這回則是30勝19敗1和。洪一中的估計, 結果是準確的。

對不同的情況, 人們常需要做估計。只是該如何估計呢? 有新聞報導, 日本九州佐賀縣有一神社, 依據稀飯發霉的情況, 及霉菌的顏色, 以估計運勢, 甚至估計何時會有地震發生。大陸揚州的算命師, 則以“耳中長毛”, 預測人可活過九十九歲 (見2006年2月號“科學人”雜誌“總編輯的話”)。如何估計, 對一般人, 可以很有道理也可毫無道理。但統計學家做估計, 總該有些依據, 有些原理。做出估計後, 更常要用以對未來做預測。統計學家的預測, 總該比用亂猜, 或諸如依稀飯、水晶球來判斷更準確才合理。量子論泰斗, 丹麥的波耳 (Niels Bohr, 1885-1962, 1922年獲諾貝爾物理獎) 曾說“預測很難, 尤其關於未來”。

世上多的是事後諸葛, 放馬後砲者。“早知道就...”, 是許多人的口頭禪。“既有今日何必當初”的懊惱, 下次仍會發生。要對過去所開出的樂透彩號碼, 找出一產生的模式是可能的, 但此模式對下一期頭獎號碼的預測, 很可能就不管用了。統計學裡有不少估計的方法, 背後其實都有一套想法在。合理的估計法, 通常也具備一些優點。但怎樣才是好的估計? 選美比賽前, 也得先規定評比標準, 否則無從選起。本文便是要對統計裡的估計, 略做討論。

二. 一葉知秋或以偏概全

民國95年1月 Yahoo 奇摩知識網站上, 有人提問: “預計三月去日本自助旅行, 估計費用大約多少?”

內容包含: 時間是五天, 地點在東京, 目的是去看世界棒球經典賽。有熱心人士回答: 台幣22,800元。算法是: 來回機票10,000元, 機場來回旅館車資2,000元, 住宿費4,000元, 生活開支4,000元, 來回球場車資800元, 購買紀念品2,000元。

回答的人可能是從他個人的經驗以做估計。他是否眞的很有經驗呢? 我們是否就可據以說, 去日本五天看球賽, 只要花22,800元? 要知回答者連球賽門票都未列入計算。

曾有一則新聞標題爲“中國醫藥界駁斥中藥致癌論 批評美國媒體以偏概全妄下結論”。以偏概全, 以管窺天, 及盲人摸象等, 都是我們批評人所見不夠客觀, 未見到全貌, 就據以下結論。但全貌常常不易見到。話又說回來, 有時全貌太大, 或太複雜, 要想見到, 可能耗時耗力耗財。像是上節中提到的全球座頭鯨的數量, 及大陸野生大熊貓數量, 大約皆無法將其全捉來數一遍。更何況就算眞見到全貌, 往往也毫無頭緒, 不知該如何下結論。不是簡單地用一句諸如“橫看成嶺側成峰”便可形容。例如, 對一初識者, 我們常只由短暫的接觸, 便對其品頭論足。但相識愈久, 看到他各方面的行爲表現, 有時反而遲疑了, 不知該如何描述他。

但我們不是也常聽到*一葉知秋*嗎? 淮南子・說山:

> 以小明大, 見一落葉而知歲之將暮; 睹瓶中之冰而知天下之寒。

這種經驗常有: 品嚐鍋中一塊肉, 就可知整鍋食物的滋味; 煮湯時, 常憑一小湯匙來決定整鍋之鹹淡; 驗血時, 也是只抽一小針筒。爲什麼這時又不會說是*以偏概全*了?

究竟何時可一葉知秋?

俗語說一樹之果有酸甜之別。但經驗告訴我們, 同一株樹所結之果, 酸甜的差異, 大致不會太大。甚至同一果園, 同一種果子的酸甜大約也差不多。淮南子一書中, 所描述的地方, 說不定每年總是到了秋天, 每株樹就開始有落葉。如果某地有不同的樹, 全年四季分別會落葉, 這時見到葉落, 就不一定是秋了。換句話說, 就是依*變異*大小, 變異小, 便容易一葉知秋, 可以嘗鼎一臠, 可以因小見大; 變異大, 即使所見再多, 仍有可能以偏概全。如果全美各地醫療不當情形大同小異, 以紐約市的資料推估全美, 就不會太離譜, 否則就是以偏概全了。

變異小的情況, 做估計當然較容易, 這時見微知著並不稀奇, 人人皆可一葉知秋。變異若不算小, 以偏便不易概全。變異大小, 深深影響估計之精準性。因此做估計時, 常須掛念變異, 常要考慮如何才能減小變異, 而這就有賴“好的”估計方法。只要變異屬於不易忽略的情況, 統計學家便有用武之地。另外, 對所欲估計的量, 若有一些事先的資訊, 也要善用, 有助於提高估計之準確性。總之, 估計人人都會, 但如何較準確的估計, 就是學問了。只是怎樣算準確呢? 有兩個錶, 一個停止不動, 永遠顯示上午6點32分, 一個每日慢1分鐘, 何者較準? 有人以爲是前者, 因它每天有一次顯示正確時間, 而後者每1,440天 (1天有1,440分), 才有一次顯示正確時間。故準確的定義爲何, 得好好斟酌。另外, 有人一葉知秋, 有人十葉知秋, 都是知秋, 顯然前者較有效率。所以除了準確性, 效率也該考慮。可看出怎樣才是好的估計, 是有不同的衡量標準。

下一節我們便來看, 對好的估計, 會有那些要求。

三. 估計之評比

假設有一銅板, 如何估計它出現正面的機率 p? 重複地投擲 n 次, 然後以總共出現之正面數除以 n, 或說以*相對頻率*(relative frequency) 來估計 p。不論你有沒有學過統計, 大約都會想到此方法。甚至這很可能是你唯一想得到的合理方法。

當然在上述估計中, 我們隱含地做了一些假設。首先每次投擲, 銅板出現正面的機率要相同, 不可以換用出現正面機率不同的銅板, 否則求出來的不知該是那一量之估計值。其次各次投擲間, 出現的結果要相互獨立。如果看到第1次得什麼面, 以後就都照抄, 如此 n 次下來, 不是 n 個正面, 就是 n 個反面。則對 p 之估計, 只會是1或0, 這樣顯然不行。

只要是可以重複觀測的試驗, 欲估計某事件發生之機率 p, 都可以採用前述相對頻率的方法。雖實驗可能有很多不同的結果, 但我們可將所有結果分成兩類, 其一是有興趣的事件 (對應銅板出現正面), 其二是該事件之餘集 (對應銅板出現反面)。觀測 n 次後, 數看看有幾次該有興趣的事件發生。在投擲銅板的例子中, 有興趣的事件, 就是銅板出現正面。這種對一參數, 以一個值來估計, 便稱點估計(point estimation)。現以 $X_1,\ldots,X_n$, 分別表各次試驗所得之結果, 其中 $X_i = 1$ 或0, 就依第 i 次試驗, 該事件發生或未發生, $i = 1,\ldots,n$。依前述討論, $X_1,\ldots,X_n$ 乃假設爲獨立且有共同分佈(independent and identically distributed, 簡稱 iid), 且 $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$。在統計裡若 $X_1,\ldots,X_n$ 爲 iid, 便稱此爲一組隨機樣本(random sample)。

有了 $X_1,\ldots,X_n$, 如何估計 p? 用樣本平均(sample mean)! 即以

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

來估計 p。而 $X_1 + \cdots + X_n$ 即爲 n 次投擲中, 共得之正面數。平均, 生活裡我們可說經常在求平均。以 $\overline{X}_n$ 估計 p 有那些優點呢? 首先因 $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$, 且

$$E(X_i) = 0 \cdot (1-p) + 1 \cdot p = p, \quad 1 \le i \le n,$$

故得

$$\begin{aligned} E(\overline{X}_n) &= E(\frac{1}{n}(X_1 + \cdots + X_n)) \\ &= \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \frac{np}{n} = p. \end{aligned}$$

即 p 之估計量 $\overline{X}_n$ 的期望值爲 p。

對一參數 θ 做估計時, 於觀測到 $X_1,\ldots,X_n$ 後, 以它們的一個函數 T_n (T_n 稱爲統計量(statistic), 用來做估計時, 又稱估計量(estimator)) 來估計 θ, 若滿足 T_n 的期望值 $E(T_n)$, 等於所欲估計的參數 θ, 即 $E(T_n) = \theta$, 則 T_n 便稱爲 θ 之不偏估計量(unbiased estimator), 簡稱爲不偏的。不偏性, 似乎是好的估計量所該具備的性質。期望值有平均的意思, 做估計有時高估, 有時低估, 但估計值總該在實際值附近打轉, 而不應傾向高估或傾向低估, 因此平均而言, 總該是準的。否則就是有偏差, 即若 $E(T_n) \ne \theta$, T_n 便稱爲 θ 之一偏差估計量。

不難看出, 前述對 p 之估計, 不偏估計量並不唯一。諸如 X_1, $(X_1+X_2)/2$, $0.2X_2+0.3X_3+0.5X_6$ 等, 皆爲 p 之不偏估計量。只要 $n\geq 2$, 便有無限多個對 p 之不偏估計量。

有些統計學家, 很在乎不偏性, 覺得好的估計量, 都至少該爲不偏的。此要求並非不合理, 只是有些不偏估計量會很荒謬。例如, 設 $X\sim\mathcal{P}(\lambda)$, $\lambda>0$, 即 X 有參數 λ 之 Poisson 分佈, 機率密度函數爲

$$f(x)=\frac{e^{-\lambda}\lambda^x}{x!},\quad x=0,1,2,\ldots.$$

現欲估計 $\theta=e^{-2\lambda}$。取統計量 $T=T(X)=(-1)^X$。則

$$E(T)=\sum_{x=0}^{\infty}(-1)^x\frac{e^{-\lambda}\lambda^x}{x!}=e^{-\lambda}\sum_{x=0}^{\infty}\frac{(-\lambda)^x}{x!}=e^{-\lambda}e^{-\lambda}=e^{-2\lambda}=\theta.$$

此處用到下述公式:

$$\sum_{x=0}^{\infty}\frac{a^x}{x!}=e^a,\quad a\in R.$$

故 T 爲 θ 之一不偏估計量。這樣的估計量有何不妥呢? 當觀測到的 X 爲偶數 $0,2,4,\ldots$ 時, 以1估計 θ, 當 X 爲奇數 $1,3,5,\ldots$ 時, 以 -1 估計 θ。但 $\theta=e^{-2\lambda}$, 其中 $\lambda>0$, 爲一介於 $0,1$ 間的數, 既不會等於1也不會等於 -1。估計量 T 雖爲不偏, 明顯地卻爲一極不合理的估計量。我們大致可以這麼說, 儘管有一些對於估計量的評比準則, 爲很多統計學家所接受, 但依這些準則, 仍有可能產生怎麼看都不像太好的估計量。

另外, 你可能也已看出, 光是要求不偏性是不夠的。對前述投擲銅板的例子, 若用 X_1 當做估計量, 即只依第一次投擲的結果來估計 p, 其他資訊 $X_2,\ldots,X_n$ 都不理會, 這樣的估計量, 顯然不會太好。直觀上, 取樣愈多 (即 n 愈大) 估計要愈準才合理。而當樣本數 n 趨近至 ∞, 更應準得不得了才對, 也就是估計量與所欲估計的參數, 差不多就該合一了。這個想法引出了一致性(consistent) 的概念。

對好的估計量之第二個要求是一致性。在此對一數列之參數 θ 的估計量 T_n, $n\geq 1$, 若滿足 $n\to\infty$ 時, T_n 機率收斂(converges in probability) 至 θ, 以

$$T_n\xrightarrow[n\to\infty]{p}\theta \tag{1}$$

表之, 便稱 T_n 爲 θ 之一致估計量(consistent estimator)。(1) 式成立的意思是

$$\lim_{n\to\infty}P(|T_n-\theta|>\varepsilon)=0,\quad \forall\varepsilon>0, \tag{2}$$

或者等價地說, 對 $\forall\varepsilon>0$ 及 $\delta>0$, 存在一 $k\geq 1$, 使得 $n\geq k$ 時,

$$P(|T_n-\theta|>\varepsilon)<\delta. \tag{3}$$

看起來有點深奧, 這是屬於機率論裡極限的題材。弱大數法則(weak law of large numbers) 告訴我們, 對一數列 iid 之隨機變數 $X_1, \ldots, X_n, n \geq 1$, 只要期望值 $E(X_i) = \mu$ 存在, 則當 $n \to \infty$ 時, 其樣本平均 $\overline{X}_n$ 會機率收斂至 μ, 即

$$\overline{X}_n \xrightarrow[n \to \infty]{p} \mu. \tag{4}$$

所以在前述投擲銅板的例子中, 由於 $\overline{X}_n \xrightarrow[n \to \infty]{p} p$, 故 $\overline{X}_n$ 為 p 之一致估計量。樣本平均 $\overline{X}_n$, 不但是 p 之不偏估計量, 且是一致估計量。至於 X_1, $(X_1 + X_2)/2$, $0.2X_1 + 0.3X_3 + 0.5X_6$ 等 p 之不偏估計量, 則非 p 之一致估計量。通常一參數 θ 之不偏估計量, 往往有無限多個, 但若再加上一致性的要求, 就會排除很多不太好的估計量。不過要討論估計量是否有一致性, 先決條件當然是樣本數 n 可無止盡的增大。

不偏性及一致性, 是我們首先會想到"好的" 估計量, 似乎該有的兩個條件。估計量仍有一些其他評比之基本準則。估計量如何才算好, 要依評比標準而定。

以一統計量 $T = T(X_1, \ldots, X_n)$ 來估計 θ (或者 θ 的一函數 $g(\theta)$), 如何衡量此估計量之表現呢? 一個最自然的度量法是看誤差 $|T - \theta|$ 之大小。但此度量法並非盡如人意。原因有二: 其一為此度量為一隨機變數, 與所觀測到的 $X_1, \ldots, X_n$ 有關, 並非一定值; 其二為此度量與未知參數 θ 有關, 因此估計到底多準, 無法得知。第一個缺點可以解決, 取期望值即可, 也就是將誤差平均掉。例如, 可以考慮平均絕對誤差 $E(|T-\theta|)$, 也可以考慮均方差(mean squared error, 簡稱 MSE) $R(\theta, T) = E((T - \theta)^2)$。此處 T 為一隨機變數, 期望值是對 T 取 (而非對 θ), 或者說對 $X_1, \ldots, X_n$ 取。以 $E(|T - \theta|)$ 或 $E((T - \theta)^2)$ 來度量誤差, 何者較適宜? 通常取後者。主要原因是, MSE 通常比 $E(|T - \theta|)$ 較易計算。又 MSE 可改寫為

$$\begin{aligned} E((T-\theta)^2) &= E((T - E(T) + E(T) - \theta)^2) \\ &= E((T - E(T))^2) + (E(T) - \theta)^2 \\ &= \mathrm{Var}(T) + b^2(\theta, T), \end{aligned} \tag{5}$$

其中 $b(\theta, T) = E(T) - \theta$。上述第二等式成立, 是用到

$$\begin{aligned} E((T - E(T))(E(T) - \theta)) &= E(T - E(T))(E(T) - \theta) \\ &= (E(T) - E(T))(E(T) - \theta) = 0, \end{aligned}$$

因 $E(E(T)) = E(T)$ (記住 $E(T)$ 為一常數)。

$b(\theta, T)$ 稱為 T 之偏差(bias)。$\mathrm{Var}(T)$ 乃用以量測估計量 T 變異之大小, $b^2(\theta, T)$ 則是量測偏差之大小。前者顯示精準性, 後者顯示正確性。比方說射飛鏢, 若射在靶上的點都很接

近, 我們可說射得很精準。但說不定這些點都很偏離紅心, 因此正確性不夠。我們實際接觸的估計量, 常都滿足$\mathrm{Var}(T) < \infty$。如果 $\mathrm{Var}(T) = \infty$, 則 $R(\theta, T) = \infty$, 這種估計量當然不好, 很少被考慮。

設有二 θ 之估計量 S, T, 所有可能的 θ 之集合, 以 Ω 表之。例如欲估計銅板出現正面的機率 θ, 則 Ω 可取成 $[0, 1]$。但如果確定 θ 只有兩個可能的值0.5及0.7, 則 Ω 可取成 $\{0.5, 0.7\}$。若 $R(\theta, T) \le R(\theta, S)$, $\forall\theta \in \Omega$, 且對某些 θ, 嚴格不等式成立, 則合理的作法是不採用 S。此時我們說 ***T* 較 *S* 爲佳**, 且稱 S *爲不可採用的*(inadmissible)。一估計量 T, 若不存在較其爲佳之估計量, 便稱爲*可採用的*(admissible)。此處所謂可不可採用, 是以 MSE 爲評比標準。評比估計量, 要留意是否爲可採用的, 此概念用途廣泛。例如, 在找對象時 (每人標準自然不盡相同), 即使未找到最佳者, 至少也要選可採用的。尺有所長, 寸有所短, 總要有某些長處, 不能樣樣皆不如人。在一團體 (如企業公司) 裡, 要儘量避免自己是一位不可採用的人。否則不但無法扮演重要角色, 當公司要裁員時, 可能也是較先被想到的。

是否有一比其他估計量都爲佳的估計量呢? 除少數特例外, 答案是否定的。假設存在一個這種估計量 T, 則任取一 $\theta_0 \in \Omega$, 且考慮估計量 $S = \theta_0$, 即恆以常數 θ_0 估計 θ。此一估計量爲高度精準 ($\mathrm{Var}(S) = 0$), 而極不正確 (除非當 θ 很接近 θ_0)。由 (5) 式, 此時 $R(\theta, S) = b^2(\theta, S)$。當 $\theta = \theta_0$ 時, $b(\theta_0, S) = E(\theta_0) - \theta_0 = 0$, 故 $R(\theta_0, S) = 0$。因 T 較 S 爲佳, 故 $R(\theta_0, T)$ 必須等於 0。但 θ_0 爲 Ω 中之任一點, 故得 $R(\theta, T) = 0, \forall\theta \in \Omega$。除了一些特殊的情況 (如 Ω 中只有一個元素), 此乃不可能。

事實上, 由上一節中, 停止的錶亦有準確時候之例, 我們早就知道, 對於估計量, 不存在一個永遠的第一名。故除了比 MSE 之大小外, 我們得再加上其他評比的準則。加上新的準則後, 可消除一些不合理的估計量 (如前述 $S(\boldsymbol{X}) = \theta_0$)。在此一包含較少估計量之集合中, 我們尋找最佳 (指 MSE 最小) 的估計量。要知理想狀況當然是*低偏差且高精準*。先控制偏差, 再考慮精準性, 似乎是合理的(想想射飛鏢的例子)。因此我們要求 $E(T(\boldsymbol{X})) = \theta, \forall\theta \in \Omega$, 或等價地說, 要求偏差等於0, 即 $b(\theta, T) = 0, \forall\theta \in \Omega$。如前所定義, 這種估計量 T, 稱爲 θ *之一不偏估計量*, 或說 T *爲不偏的*。若 T 爲 θ 之一不偏估計量, 則 $R(\theta, T) = \mathrm{Var}(T(\boldsymbol{X}))$。

要求不偏除了合理外, 常可在所有不偏估計量中, 找到一較所有其他估計量“不差”的估計量。即有一估計量 T^*, 滿足 $E(T^*) = \theta$, 且對任一其它不偏估計量 S, 皆有

$$R(\theta, T^*) = \mathrm{Var}(T^*) \le \mathrm{Var}(S) = R(\theta, S),\ \forall\theta \in \Omega. \tag{6}$$

這種 T^* 稱爲*一致最小變異不偏估計量*(uniformly minimum variance unbiased estimator, 簡稱 UMVUE)。所謂“一致”, 是指對每一 $\theta \in \Omega$, T^* 之變異數皆最小。

數理統計中, 會給出一些有效的步驟, 以找出參數之 UMVUE, 此處不多討論。不過必須一提的是, 有時不偏估計量不存在, 因此也就沒有 UMVUE。有時 UMVUE 雖存在, 但可能

是很荒謬的估計量, 或有時是不可採用的估計量, 或有時計算上很複雜。前述對 $\mathcal{P}(\lambda)$ 分佈, 估計 $\theta = e^{-2\lambda}$ 的例子裡, $T = (-1)^X$ 雖爲一極不合理的估計量, 但可以證明, 它是 θ 唯一之不偏估計量, 因此當然也就是 θ 之 UMVUE。在只有一個估計量的集合裡, 找最佳者, 不論是在那一種意義下的最佳, 都不會是太令人樂於接受的估計量。一般而言, 並沒有那一估計量是全方位最佳的。我們在所有不偏估計量的集合中找*最佳估計量*, 而所謂最佳乃指變異數最小。僅是依不偏及變異數最小來評比, 通常還合理。但若找到的最佳估計量不盡如人意, 也不用太奇怪。此正如選美不論評比標準爲何, 選出來的第一名, 不見得人人叫好。

UMVUE是在樣本數 n 固定下找的。如果樣本數 n 可以無止盡的增加, 估計量的近似行爲, 便也可拿來評比。我們期望在樣本數很大下, 估計量會有一些好的性質。付出了較大的代價(取樣增大), 總該有一些所得。這方面的討論就是統計裡的*大樣本理論*(large sample theory)。之前已給出一致估計量, 另外還有*漸近不偏估計量*(asymptotic unbiased estimator)。即一數列 θ 之估計量 $T_n,\ n \geq 1$, 滿足

$$\lim_{n\to\infty} E(T_n) = \theta. \tag{7}$$

如果對 $\forall n \geq 1$, T_n 皆爲 θ 之不偏估計量, 即 $E(T_n) = \theta,\ \forall n \geq 1$, 則 T_n 當然也是 θ 之漸近不偏估計量。但如果 T_n 不是不偏估計量, 退而求其次, 當樣本數很大時, 若 T_n 就"差不多"是不偏的, 即 (7) 式成立, 那便也不錯。其他還有*漸近相對有效性*(asymptotic relative efficiency)、漸近有效估計量、漸近常態等, 皆是大樣本下, "好的"估計量可能會有的性質。

設有 $T_1 = \{T_{1n}, n \geq 1\}$ 及 $T_2 = \{T_{2n}, n \geq 1\}$ 二數列之不偏估計量。又設 $\mathrm{Var}(T_{1n}) = (n+100)/n^2,\ \mathrm{Var}(T_{2n}) = 2n/n^2,\ n \geq 1$。則當 $n = 1, 2, \ldots, 99$, 皆有 $\mathrm{Var}(T_{1n}) > \mathrm{Var}(T_{2n})$, 此時 T_{2n} 較佳; $n = 100$ 時, $\mathrm{Var}(T_{1n}) = \mathrm{Var}(T_{2n})$; 但自 $n = 101$ 起, 便皆有 $\mathrm{Var}(T_{1n}) < \mathrm{Var}(T_{2n})$, 且

$$\lim_{n\to\infty} \frac{\mathrm{Var}(T_{1n})}{\mathrm{Var}(T_{2n})} = \lim_{n\to\infty} \frac{1/n^2}{2/n^2} = \frac{1}{2} < 1.$$

當 $\lim_{n\to\infty} \mathrm{Var}(T_{1n})/\mathrm{Var}(T_{2n}) < 1$, 我們便稱 T_1 比 T_2 *漸近有效*。表 n 夠大時, 欲達到同樣的精準度, 估計量 T_2 所需之樣本數比 T_1 多。樣本的取得有時是很不容易的, 從取樣之角度而言, 在大樣本下, T_1 比 T_2 有效率。另外, 設

$$\mathrm{Var}(T_{1n}) = \frac{n^2}{(n+1)^2(n+2)},\ \mathrm{Var}(T_{2n}) = \frac{1}{n},\ n \geq 1,$$

則雖 $\mathrm{Var}(T_{1n}) \leq \mathrm{Var}(T_{2n}),\ \forall n \geq 1$, 但因 $\lim_{n\to\infty} \mathrm{Var}(T_{1n})/\mathrm{Var}(T_{2n}) = 1$, 故定義 T_1, T_2 之漸近有效性相同。

大樣本理論是統計裡重要的題材。從微積分開始, 各種問題中, 極限下的性質常是大家感興趣的。有關估計量評比的介紹, 我們就此打住。欲進一步了解的讀者, 可參考一般數理統計的書。

四. 估計的方法

估計要準確容易嗎? 不知大家是否有下述這類經驗: 去看電影, 或是去餐廳, 經常是人滿爲患, 看來景氣極佳。但電影業, 餐飲業卻常抱怨生意不好做。學生修通識課程, 常選不上, 似乎班班爆滿。但實際上卻有很多通識課程, 修課人數並不太多。這究竟是怎麼回事。

假設有家餐廳, 有100個座位。週一至週五, 每晚只有10個人去用餐, 而週六及週日, 兩晚皆客滿, 若未預訂是沒有座位的。一週7個晚上共250個人去用餐, 則平均每晚只有 $250/7 \doteq 35.71$ 個客人, 生意並不算太好。但有80%(200/250) 的顧客, 去用餐時 (週末), 見到餐廳座無虛席, 他們的印象是餐廳生意眞好。再看一例。假設有5班通識課程, 1班有100個人修, 另4班各只有10人修。5班修課人數共 $100+4\cdot 10=140$ (人), 平均每班只有 $140/5=28$ (人)。但卻有 $100/140 \doteq 71.43\%$ 的學生, 上課得提早去搶位子。現對這5班全部140個學生發問卷, 調查其修課的班上有多少人? 又假設此140張問卷皆收回, 則其中有100張塡100人, 40張塡10人, 總共所得修課人數有 $100\cdot 100+40\cdot 10=10,400$ (人)。問卷顯示, 平均每班有 $10,400/140 \doteq 74.28$ (人)。

這類例子很多, 旅遊名勝景點, 遊客常抱怨這麼賺錢, 卻不提高服務品質。連餐廳都不多設幾處, 財源滾滾卻不想讓它們進來, 停車位也太少。眞實的情況是, 除了少數假日遊客摩肩接踵外, 其餘的日子, 大多是門可羅雀。在上述這種情況裡, 若遊客 (或學生) 由其親身體驗, 以估計該景點平均每天有多少遊客 (或每門課有多少人修), 大部分人的估計, 都可能極不正確。

人類是習於觀察的。觀察到天體運行的規律性, 四時的變化, 動物及植物的生長模式。我們做決策 (或做估計), 往往也是先觀測, 再依所得的資料(data, 或說數據), 而做出推論。沒有資料, 統計學家是束手無策的。在 Conan Doyle 所著的福爾摩斯全集(Sherlock Holmes stories), The Adventure of the Copper Beeches 一書中, 寫著:

"Data! Data! Data!" he cried impatiently.

"I can't make bricks without clay."

福爾摩斯說沒有 data 無法做判斷, 正如沒有黏土 (clay) 無法做磚 (brick)。但資料若有問題, 所得的推論當然不會好。像是調查局如果情資的來源爲媒體報導, 或政論節目內容, 可信度當然不高。資料的取得及整理, 在統計裡是一大學問, 本文不擬探討。假設資料沒問題, 則有那些估計的方法呢? 這要視估計的對象及不同的情況而定, 本節我們仍僅就參數的點估計討論。

業餘球類比賽, 單淘汰或雙淘汰是常採的方法。實力雖不錯, 但運氣不好, 包括球員臨場表現不佳, 裁判誤判, 或賽程排得不利, 可能早早被淘汰。職業球賽如籃球、棒球等之比賽, 則是依整個球季之勝率, 以定出那些隊可晉級。一個球季比賽場數很多, 像是職棒比賽, 一個球季少說有百場, 美國大聯盟每隊更要打滿162場。計算勝率, 大致可稀釋掉運氣等因素, 以呈現每一球隊的眞正實力。在上一節中, 我們以樣本平均 $\overline{X}_n$, 估計 X_i 的期望值 $E(X_i)$。表面上看起來, 是因期望值有平均的意思, 背後的理論依據爲*大數法則*(law of large numbers)。大數法則有兩個版本, 比較簡單的是上一節所介紹的弱大數法則, 另一爲*強大數法則*(strong law of large numbers)。

由於當 $X_1, \ldots, X_n$, $n \geq 1$, 爲 iid, 且 $E(X_i^2)$ 存在時, 弱大數法則告訴我們, $X_1^2, \ldots, X_n^2$ 的樣本平均 $\overline{X_n^2}$, 機率收斂至 $E(X_i^2)$。即

$$\overline{X_n^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow[n\to\infty]{p} E(X_i^2). \tag{8}$$

所以若欲估計參數 $E(X_i^2)$, 可以 $\overline{X_n^2}$ 做爲估計量。以平均估計期望值, 這樣所獲的估計量, 通常還不太壞, 至少爲不偏且有一致性。

例如, 設 X 有常態分佈, 期望值爲 0, 變異數爲 σ^2, 即 $X \sim \mathcal{N}(0, \sigma^2)$。因 $E(X^2) = \sigma^2$, 所以若欲估計 σ^2, 且觀測到隨機樣本 $X_1, \ldots, X_n$, 則可以 $\sum_{i=1}^{n} X_i^2/n$ 做爲估計量。由於對任一整數 r, X^r 的期望值 $E(X^r)$ 稱爲 X 之 r 次動差。所以上述這種經由欲估計的參數, 是 X 的幾次方的期望值, 便用對應的 X 之次方的樣本平均來估計, 便稱*動差法*(method of moments), 所得之估計量, 稱爲*動差估計量*(moment estimator)。動差估計量並不唯一。例如, 設 $X \sim \mathcal{P}(\lambda)$。因 $E(X) = \mathrm{Var}(X) = \lambda$, 而 $\mathrm{Var}(X) = E(X^2) - (E(X))^2$, 故

$$\overline{X}_n \quad \text{及} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X}_n)^2$$

皆可用來當做 λ 之動差估計量。

這種依相對頻率來做估計, 稱爲*古典的作法*(classical approach), 採這種主張的統計學家, 稱爲*頻率學派*(frequentist)。在十九世紀, 這派想法可說是主導統計的應用。直至今日, *頻率的觀點*, 仍是一很基本的統計思維。

底下介紹另一種估計法。

教室玻璃被打破了, 老師從平常最調皮的同學開始問。有命案發生, 從現場採到的指紋開始追查。皆是因認爲這些人嫌疑最大。醫生看診, 常也是從病人的症狀, 推測那一種病最易產生此症狀。從所得之觀測值, 推測究竟參數爲何, 會使得到此一觀測值之機率最大, 這也是一種常用的估計方法。在統計學裡稱爲*最大概似法*(method of maximum likelihood)。所得之估計

量稱爲最大概似估計量(maximum likelihood estimator, 簡稱 MLE)。這種估計法有其道理, 但會不會誤判? 當然會。只是警方辦案, 不從有前科、有地緣關係、由現場蒐到的可疑事物開始追查, 難道要從毫不相干者開始查? 那不是更不合理嗎? 統計理論顯示, 最大概似估計量, 有很多好的性質。例如, 常也就是 UMVUE, 常也有漸近常態分佈等。因此這也是廣爲被採用的一種估計法。

例 1: 設一盒子中有不少白球及黑球, 假設只知數目比爲3比1, 但不知是白球多還是黑球多。依序隨機地抽取3球, 每次取出後放回。令 X 表所得之黑球數。則由假設知 X 有 $\mathcal{B}(3,p)$ 分佈, 其中 $p=1/4$ 或 $3/4$, 即 X 之 p.d.f. 爲

$$f(x|p)=P(X=x)=\binom{3}{x}p^x(1-p)^{3-x},\ x=0,1,2,3,$$

且 $p=1/4$ 或 $3/4$。

我們想根據觀測到之黑球數 x 來估計 p。此估計並不算困難, 因只有二選擇: $1/4$ 或 $3/4$。抽取後, 有四種可能的結果, 其機率值爲

x	0	1	2	3
$f(x\|1/4)$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$
$f(x\|3/4)$	$\frac{1}{64}$	$\frac{9}{64}$	$\frac{27}{64}$	$\frac{27}{64}$

如果 $x=0$, 就宜選 $p=1/4$, 因爲 $27/64$ 的機率值較 $1/64$ 爲大。也就是 $p=1/4$ 比 $p=3/4$, 更會使 $x=0$ 發生。所以 $x=0$ 出現時, 認爲 $p=1/4$ 才較合理。$x=1$ 時, 也宜選 $p=1/4$。至於若出現 $x=2$ 或 3, 則 $p=3/4$ 當然是較合理的選擇。故 p 之最大概似估計值 $\hat{p}$ 爲

$$\hat{p}=\begin{cases}1/4, & \text{若} x=0,1,\\ 3/4, & \text{若} x=2,3.\end{cases}$$

對同一參數所做之估計, 最大概似法與動差法, 有時會得到相同的估計量。這並不足爲奇, 因二者皆是很好的估計法, 難免英雄所見略同。不過因畢竟是兩種不同的方法, 當然也有所得之估計量不同的時候。

除了以上所介紹的兩種估計法外, 另有一種常用的估計法, 就是貝氏作法(Bayesian approach), 持此主張者, 也稱貝氏學派(Bayesian)。在前兩種估計法中, 欲估計的參數 θ, 被視爲一未知但固定的值。但在有些情況, 我們對 θ 有一些事先的了解。例如, 設 θ 爲銅板出現正面的機率。有人以爲, 由於這是政府發行的銅板, 不至於太離譜, θ 應在0.5附近, 甚至認爲 θ 就

在0.45至0.55間, 且在這區間中的所有值, 可能性皆相同。因此假設 θ 在區間 $[0.45, 0.55]$ 均勻分佈。即 θ 有 $\mathcal{U}[0.45, 0.55]$ 分佈。

對 θ 假設有一事前分佈(prior distribution)。此分佈可視爲主觀的分佈(subjective distribution), 因此乃做估計者主觀的看法。當然主觀的看法, 也可以是依據過去的資料, 而做出客觀的推斷。要知或出於天生, 或因擁有的資訊很豐富, 有些人主觀的看法可能很客觀。而合理的思維是, 於取樣後, 根據所得之觀測結果, 修正對 θ 之分佈的看法, 如此便得到事後分佈(posterior distribution)。貝氏此一名稱的由來, 是因這一經修正所得的分佈, 是利用貝氏定理(Bayes' rule) 而得。貝氏(Thomas Bayes, 1702-1761) 爲英國牧師, 但其姓氏早已成爲一重要的統計名詞。

事後分佈是一條件分佈, 給定的條件是觀測到的樣本。利用事後分佈, 以對 θ 做推估。本來我們對 θ 有一看法 (事前分佈), 看到樣本後, 看法改變 (事後分佈), 基於此新看法以估計 θ。事後分佈與事前分佈往往有很大差異。可以一簡單的例子來說明。有人敲門, 若沒有其他資訊, 則可合理地假設是男或是女的機率, 各爲 $1/2$。這是事前機率。但若由虛掩的門, 看到來者穿裙子, 這資訊就要好好利用。男生穿裙子的很少見吧, 大約只有蘇格蘭人, 假設任一男生穿裙子的機率爲 $1/10,000$。女生也不見得都穿裙子, 假設任一女生穿裙子的機率爲 $3/10$。所以任一敲門者, 穿裙子的機率爲

$$\frac{1}{10,000}\cdot\frac{1}{2}+\frac{3}{10}\cdot\frac{1}{2}=\frac{3,001}{20,000}.$$

因此給定敲門者穿裙子的條件下, 來者是女生的機率爲

$$P(\text{女生}|\text{敲門者穿裙子})=\frac{P(\text{敲門者爲女生且穿裙子})}{P(\text{敲門者穿裙子})}$$

$$=\frac{\frac{3}{10}\cdot\frac{1}{2}}{\frac{3,001}{20,000}}=\frac{3,000}{3,001}\doteq 0.9996667777.$$

即在我們的假設下, 當觀測到敲門者“穿裙子”, 則敲門者是女生的機率, 由事先的 $1/2$, 增高至一很接近1的值。至於是男生的機率, 則從 $1/2$ 降至很接近0的 $1/3,001$。事後分佈與事前分佈, 可能會有極大不同, 貝氏學派認爲不容忽視此點。他們主張事先掌握的資訊 (事先分佈), 就是該充分利用, 而也必須依觀測值做修正 (以到事後分佈), 再依據事後分佈, 進行推估。

註 1: $3,000/3,001$ 並非就是一直 0.9996667777 循環下去, 接著的幾位爲 40753 等。事實上, 若將此數展開爲小數, 其循環節長達 $1,500$ 位。

一個常見的 θ 之貝氏估計量, 爲給定觀測值 X 下, 事後分佈的期望值, 即 $E(\theta|X)$。我們稱此爲貝氏估計量(Bayesian estimator), 不過也還有很多其他不同的貝氏估計量。近年來, 貝氏的作法愈來愈被廣爲採用, 特別是在醫學及商業上。

上述這三種估計法, 反映三種不同的統計估計思維。在做估計時, 有時是三種方式混合著用。視不同的情況或所擁有不同的資訊而定。要善用這三種方法, 以得到較好的估計品質。

對於一參數 θ, 除了點估計外, 還有區間估計(interval estimation)。即給出一隨機區間, 並指明此爲 θ 之若干百分比 (通常取90%、95%或99%等) 的信賴區間(confidence interval)。信賴區間的討論, 可參考黃文璋 (2006) 一文。

不只是估計一參數 θ, 有時可估計 θ 的某一函數 $g(\theta)$, 或二參數 θ_1, θ_2 的某一函數 $h(\theta_1, \theta_2)$ 等。例如, 對 $\mathcal{N}(\mu, \sigma^2)$ 分佈, 估計 μ/σ。這些都是有關參數估計的問題, 其他尚有分佈的估計等。估計論是統計裡一重要的題材, 這方面的探討很多。本文爲一初步的介紹, 無法多加討論。對估計論有興趣的讀者, 可參考一般數理統計的書。

五. 迴歸效應

本節我們介紹一種並非參數估計的估計法。

迴歸一詞, 爲十九世紀英國統計學家高頓 (Francis Galton, 1822-1911) 所首先引進, 以描述諸如父親身高與兒子身高等二變數間的關係。設以 x_i 表第 i 個父親的身高, Y_i 表其兒子之身高, $i = 1, \ldots, n$。高頓想建立一如下式之 x 與 Y 的關係:

$$Y = Q(x) + \varepsilon, \tag{9}$$

即以父親的身高 x, 來預測兒子的身高 Y。爲何父親身高以小寫, 兒子身高以大寫? 對那些父親身高爲172公分者, 高頓並不將172視爲隨機變數。但同是172公分高的父親, 兒子則不一定有多高, 所以將兒子身高 Y 視爲一隨機變數。在 (9) 式中, 隨機變數 Y, 可表示爲一 x 的函數 $Q(x)$ (不爲隨機的量), 加上一項隨機的誤差 ε。由所收集到的數據 $(x_1, y_1), \ldots, (x_n, y_n)$, 來估計函數 $Q(x)$。身高爲 x 的父親, 其小孩的身高便爲 $Q(x) + \varepsilon$。此雖是一隨機變數, 但若誤差項 ε 不太大, 則以 $Q(x)$ 來估計 Y, 就不會太離譜。比較一般的模式是, 對不同的 x, 誤差可能不同。所以可考慮下述模式:

$$Y_i = Q(x_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{10}$$

迴歸分析(regression analysis), 特別是線性迴歸(linear regression), 幾乎可說是一最常用的統計方法。有各種形式的迴歸, 如線性、非線性、多元、參數 (parametric) 及非參數 (non-parametric) 等。這其中最簡單的, 應屬簡單線性迴歸(simple linear regression), 即 $Q(x) = \alpha + \beta x$, 而 α, β 爲二常數。稱"簡單"是爲了與多元線性迴歸 (multiple linear regression) 區隔。

如前所述, 迴歸分析是爲了探討一變數與另一變數相依的情形。設有一簡單線性迴歸模式, 即設

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{11}$$

其中 α, β 爲二未知常數。又 $\varepsilon_1, \ldots, \varepsilon_n$ 常設爲獨立, 但分佈不一定要相同。有時僅假設 $\varepsilon_1, \ldots,$ ε_n 爲無相關, 即 $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0,\ 1 \le i < j \le n$。統計模型總是可以有各種假設, 只要看合不合理, 及是否符合實際情況。對 $\varepsilon_1, \ldots, \varepsilon_n$, 通常又假設期望值皆爲0, 變異數亦皆相同, 即設 $E(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma^2,\ i = 1, \ldots, n$。此時由 (11) 式得

$$E(Y_i) = \alpha + \beta x_i. \tag{12}$$

在迴歸分析裡, 當知道 x_i 之值, 我們常以 (12) 式右側, 做爲 Y_i 之估計值。(12) 式給出 $E(Y_i)$ 爲 x_i 的函數, 稱爲*母體迴歸函數*(population regression function)。至於 α, β 如何估計呢? 一個常用的方法是*最小平方法*(least squares method), 此法可追溯到法國數學家*拉格朗治*(Joseph Louis Lagrange, 1736-1813), 及德國的*高斯*(Carl Friedrich Gauss, 1777-1855)。對於如 (11) 式之模式, 於觀測到 $(x_1, y_1), \ldots, (x_n, y_n)$ 後, 最小平方法乃基於使誤差平方和

$$U(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{13}$$

達到最小值的原理, 以求出 α, β。將函數 $U(\alpha, \beta)$, 分別對 α, β 微分, 且皆令其爲0, 得

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0, \qquad \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0.$$

由上二式解出 α, β 之估計值

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{14}$$

$$\hat{\alpha} = \bar{y} - \beta \bar{x}, \tag{15}$$

其中

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

由上述方法所得之 α, β 之估計量, 便稱爲*最小平方估計量*(least squares estimators)。我們便以 $\hat{\alpha} + \hat{\beta} x_i$ 做爲 Y_i 之估計。

線性函數當然是一極簡單的函數。簡單線性模型, 就是以一線性函數 $\alpha+\beta x$, 由知道 x, 來預測 Y。這種作法, 不用知道 X, Y 二變數之聯合分佈, 樣本數 n 也不用很大。如果將所觀測到之 $(x_1, y_1), \ldots, (x_n, y_n)$, 描繪於 $x-y$ 座標平面上, 所得的直線 $y=\hat{\alpha}+\hat{\beta}x$, 便是在最小平方法的意義下, 最能描述這些數據的一條直線。底下給一例子。

例 2: McDonald and Studden(1990) 曾研究汽車排出碳氫化合物 (hydro- carbon) 的量。他們得到下述成對總哩程數 x (單位爲千英哩), 與碳氫化合物 y (單位爲公克/英哩) 的值如下:

$$\begin{aligned}&(5.133, 0.265), (10.124, 0.278), (15.060, 0.282), (19.946, 0.286),\\&(29.792, 0.333), (29.877, 0.343), (35.011, 0.335), (39.878, 0.311),\\&(24.899, 0.310), (44.862, 0.345), (49.795, 0.319).\end{aligned}$$

在此 $n=11$, 且 $\sum_{i=1}^{11} x_i = 304.377$, $\sum_{i=1}^{11} x_i^2 \doteq 10461.814$, $\sum_{i=1}^{11} y_i = 3.407$, $\sum_{i=1}^{11} y_i^2 \doteq 1.063$, $\sum_{i=1}^{11} x_i y_i \doteq 97.506$。故得 $\overline{x} \doteq 27.671$, $\overline{y} \doteq 0.310$, 且

$$\begin{aligned}&\hat{\beta} \doteq \frac{97.506-304.377\cdot 3.407/11}{10461.814-(304.377)^2/11} \doteq 0.00158,\\&\hat{\alpha} \doteq 0.310-0.00158\cdot 27.671 \doteq 0.266.\end{aligned}$$

即可以簡單線性迴歸函數 $y=0.266+0.00158x$ 來預測 y。例如, 若汽車行駛 30,000 英哩, 則預測 $y=0.266+0.00158\cdot 30=0.3134$(公克)。

我們繪出 (x_i, y_i), $i=1, \ldots, 11$, 及估計的簡單線性迴歸函數於圖1。

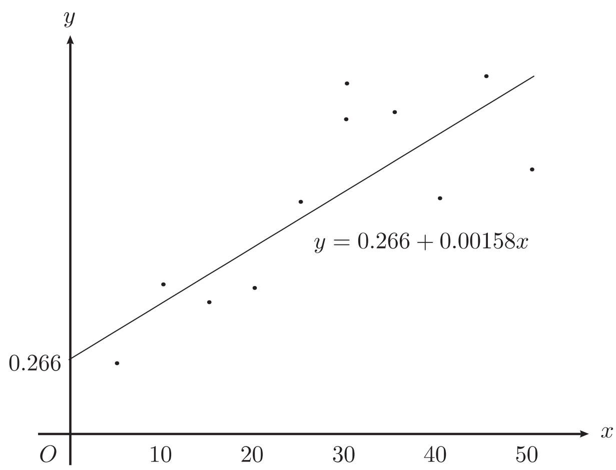

圖1. 簡單線性迴歸函數, x 軸單位爲千英哩, y 軸單位爲公克/英哩。

高頓在研究父親與兒子身高的關係時, 發現較高的父親, 兒子有較高的傾向, 而較矮的父親, 兒子亦有較矮的傾向。這當然是衆所周知, 並非一了不起的發現, 不過就是遺傳而已。但他也發現那些較高的父親, 兒子往往卻稍矮些; 而很矮的父親, 兒子卻常稍高些。此現象即爲著名的迴歸效應(regression effect), 高頓稱此現象爲向平均迴歸(regression toward the mean)。regression 有回歸, 後退, 或退化的意思。在此即表靠向平均。大致可以這麼說, 迴歸的意思就是回歸。就如我們常說的回歸正傳, 回歸祖國, 回歸專業倫理等。而迴歸效應是向平均回歸。這是迴歸這一名詞的由來。由於迴歸與回歸意思近似, 因此也有人使用回歸及回歸分析。

在高頓所獲得之1,078對父子的身高, 父親的平均身高約爲68英吋, 兒子的平均身高約爲69英吋, 高了一英吋。樣本標準差則皆約爲2.5英吋。對一72英吋高之父親, 我們會猜想其兒子應有73英吋高, 而對一64英吋高之父親, 會猜想其兒子應有65英吋。眞實的情況是, 72英吋高的父親, 兒子之平均身高約爲71英吋, 而64英吋高之父親, 兒子之平均身高約爲67英吋, 皆向平均69英吋靠近。

其實不獨父子的身高, 考試成績亦有此現象。計算兩次考試成績, 假設兩次全班的平均成績差不多。但第一次成績較高的那群學生, 第二次的成績常略降低, 而第一次成績較低的那群學生, 第二次的成績常略提高。迴歸效應再度發威。

是老天爺有心做一些平衡嗎? 倒也非如此。以身高爲例, 來解釋實際上的確會這樣。假設某地區30歲男子的身高, 最高者爲240公分。不需再高些, 就假設這群人兒子的平均身高亦爲240公分。因不致於全部兒子都等高, 故其中必會有比240公分還矮者, 也會有比240公分還高者。如此一代超越一代, 人類身高之極大値將不斷提高, 但情況顯然並非如此。故身高240公分的那群父親, 其兒子的平均身高應小於240公分。對身高最矮的那群父親, 其兒子的平均身高, 也有類似的解釋。

再以考試爲例。考試成績爲眞實的成績加上誤差。第一次考高分者, 運氣總不會是壞的。第二次就不容易在好運之上再加上好運。例如, 第一次考98分的人, 往上只有兩分的成長空間, 要提高相當不容易。而第一次只考3分的人, 第二次又能再怎麼往下掉呢? 往上升還較可能。這樣看來, 迴歸效應的存在, 就不是太怪異了。早期有些中小學老師以嚴厲著稱。對考試成績不佳的學生, 常予以處罰, 而對考得好的學生, 也少有讚美。他們是不相信鼓勵這一套的。因長期的經驗告訴他們, 經過處罰後, 原先考差的學生, 有不少下次的確進步了。至於原先考好的學生, 下次卻常略有退步, 若當初還讚美他們, 不是讓他們更鬆懈, 退步更多? 因此那些老師堅信其作法是對的。你現在知道了, 這些成績的進步或退步, 很可能只是迴歸效應, 而非老師的作法生效。

六. 結語

我們經常在做估計, 生活上、商業上、科學上、政治上, 各種估計不斷地在進行。讀者是否注意到 “做” 這一個字的中間及左邊即爲 “估”。各種估計法, 背後都有些原理, 也反映人們做決

策的某種思維。沒有一種估計法, 能行走天下萬無一失。本文只是對統計裡的估計, 略做介紹。讓讀者得到一些基本的概念。有興趣的讀者, 可從汗牛充棟的統計書中, 找到各種統計估計的方法。

参考文獻

1. 黃文璋 (2006). 統計裡的信賴。數學傳播季刊, 30卷4期 (120), 民95年12月。
2. McDonald, G.C. and Studden, W.J.(1990). Design aspects of regression-based ratio estimation. *Technometrics*, 32, 417-424.

—本文作者任教於國立高雄大學應用數學系—