信賴區間與假設檢定

黃文璋

國立高雄大學應用數學系

1. 前言

投擲一銅板 k 次, 假設銅板各次出現的結果為獨立, 且每次出現正面的機率皆為 p, 則總共得到的正面數 X 有二項分佈(binomial distribution), 參數分別為 k 及 p, 我們以 $\mathcal{B}(k,p)$ 表此分佈。即 X 會是 i 的機率為

$$P(X=i)=\binom{k}{i}p^i(1-p)^{k-i}, i=0,1,\cdots,k\ 。\tag{1}$$

投擲一銅板 k 次, 會得到幾次正面是不一定的, 除非此銅板兩面皆為正(此時 $p=1$), 或兩面皆為反(此時 $p=0$), 否則所得的正面數 X, 其值從 $0,1$ 至 k 皆有可能。此種現象稱為一隨機現象(random phenomenon), X 則稱為一隨機變數(random variable)。由排列組合中所學到的技巧, 我們可輕易得到(1)式。也就是對投擲銅板會得到幾個正面此一隨機現象, 由理論上的結構, 我們給出了(1)式為其機率模型(probability model)。其中有一參數(parameter)p 則仍屬未知, 除非知道 p 之值, 否則(1)式中之機率還是求不出的。

在量測時, 我們常假設誤差為常態分佈(normal distribution), 參數分別為 μ 及 σ^2, 以 $\mathcal{N}(\mu,\sigma^2)$ 表之。若再加上誤差對稱於 0的假設, 則 μ 便取為 0。人的身高、體重及智商等, 便是常以常態分佈為其機率模型的例子。這其中主要的理論依據則為著名的中央極限定理(Central limit theorem)。

其他還有許多分佈是常被拿來當作機率模型的, 大家可翻閱一般機率論的書, 在此不多介紹。一隨機現象(或一統計實驗(statistical experiment))之所有的結果, 便構成所謂的母體(population)。如前述投擲銅板的例子, 若做了 n 次實驗, 且以 $X_1,X_2,\cdots,X_n$ 分別表此 n 次所得之正面數, 則 $X_1,X_2,\cdots,X_n$ 為獨立且有共同分佈(independent and identically distributed, 簡稱i.i.d.), 稱為自機率密度函數(probability density function, 簡稱p.d.f.) 如(1)式之母體所產生之一組隨機樣本(random sample), 簡稱樣本。

一旦一機率模型被建立了, 通常便要估計(estimate) 其中未知的參數。統計學裏便發展出許多估計的方法(method of estimation)。而所謂點估計(point estimator), 便是一組樣本 $X_1,X_2,\cdots,X_n$ 之某一函數 $W(X_1,X_2,\cdots,X_n)$。而 $W(X_1,X_2,\cdots,X_n)$ 又稱為一統計

量(statistic)。因此任一統計量皆可當作一點估計。又為了判斷這些不同的估計方法孰優孰劣? 便發展出評估估計量(estimator)的方法(method of evaluating estimators)。

在一機率模型中, 設以0.3來估計其中某一參數 θ, 則"$\theta = 0.3$"此一敘述, 便稱為一假設(hypothesis, 複數為hypotheses)。所謂假設就是對母體中之參數的一些"看法", 或說一判斷。此假設是要經過檢定(test), 才知其正確與否。檢定一假設的過程, 便稱假設檢定(test of hypothesis, 或hypothesis testing)。一般所謂統計推論(statistical inference), 便是包含估計及假設檢定兩部份。

對一參數 θ,點估計給出此參數之一估計值。不過有時需要知道估計的可靠程度, 這時便要給出一個區間, 並且指出此區間包含 θ 之機率, 這就是區間估計 (interval estimation)。給定一 $0 < \alpha < 1$, 若存在二統計量 $U = U(X_1, X_2, \cdots, X_n)$, 及 $L = L(X_1, X_2, \cdots, X_n)$,$L \leq U$, 並滿足

$$P(L \leq \theta \leq U) = 1 - \alpha, \tag{2}$$

則隨機區間 $[L, U]$ 稱為 θ 的一 $100(1-\alpha)\%$ 信賴區間(confidence interval), 至於 $1-\alpha$ 則稱為信賴係數(confidence coefficient), 或稱信賴度, L 與 U 則分別成為信賴下界及信賴上界。信賴區間又稱為置信區間, 信賴係數又稱置信度。

以一區間估計來取代點估計的目的, 就是為使對參數的掌握能有一些保證。例如, 若估計銅板出現正面的機率 $p = 0.4$, 則因 $p \in [0, 1]$, 為一連續的區間, 故此估計會命中實際的 p 之可能性大約是零。但若給出一區間, 譬如說 $[0.3, 0.5]$, 且指出 p 會落在此區間的機率為 0.95, 則對 p 之大小反而有一更清晰的概念。這是為什麼有時要討論區間估計, 而不僅是點估計的主因。

有時候特別是對離散型的分佈, 不一定可找到一區間 $[L, U]$, 使得 $P(L \leq \theta \leq U)$ 剛好等於 $1 - \alpha$。此時我們便尋找 L 及 U, 使得 $P(L \leq \theta \leq U) \geq 1 - \alpha$, 且使左側機率儘可能地接近 $1 - \alpha$。

信賴區間與假設檢定雖為統計學中兩個基本的題材, 其原理也都不太難, 但有些人在使用時, 因不夠謹慎常犯了錯誤而不自知。本文便是要針對此二題材, 略加闡釋。我們並不擬完整地介紹此二題材, 因這是一般統計學教科書的工作, 我們只是要使各位因此能在日後隨時留意, 務必要正確而又有效地使用統計方法。

2. 信賴區間

設隨機變數 $X_1, X_2, \cdots, X_n$ 為i.i.d. 以 $\mathcal{N}(\mu, \sigma^2)$ 為其共同分佈, 令 $\overline{X}_n = \sum_{i=1}^n x_i/n$ 表其樣本平均。則知 $\overline{X}_n$ 有 $\mathcal{N}(\mu, \sigma^2/n)$ 分佈。令 Z 表一有 $\mathcal{N}(0, 1)$ 分佈(即標準常態分

佈(standard normal distribution))之隨機變數, 且令 Φ 表其分佈函數(distribution function), 即

$$\Phi(x) = P(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du, \quad x \in R,$$

又令 z_y 表 $\Phi(x)$ 之反函數, 即對 $\forall\ 0 < y < 1$,

$$\Phi(z_y) = P(Z \leq z_y) = \int_{-\infty}^{z_y} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = y\ 。$$

現因對 $\forall\ 0 < \alpha < 1$,

$$\begin{aligned} P(& \overline{X}_n - \sigma z_{1-\alpha/2}/\sqrt{n} \leq \mu \leq \overline{X}_n + \sigma z_{1-\alpha/2}/\sqrt{n}) \\ = & \ P\left(\left|\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right| \leq z_{1-\alpha/2}\right) \\ = & \ P(|Z| \leq z_{1-\alpha/2}) = 1 - \alpha/2 - \alpha/2 = 1 - \alpha, \end{aligned}$$

即

$$P(\overline{X}_n - \sigma z_{1-\alpha/2}/\sqrt{n} \leq \mu \leq \overline{X}_n + \sigma z_{1-\alpha/2}/\sqrt{n}) = 1 - \alpha\ 。 \tag{3}$$

故若 σ 已知, 則

$$I = [\ \overline{X}_n - \sigma z_{1-\alpha/2}/\sqrt{n}\ ,\ \overline{X}_n + \sigma z_{1-\alpha/2}/\sqrt{n}\ ] \tag{4}$$

為 μ 之一信賴係數為 $1-\alpha$ 之信賴區間。此區間有時以 $I = \overline{X}_n \pm \sigma z_{1-\alpha/2}/\sqrt{n}$ 表之。給定一 α 值, 由標準常態分佈之數值表可查出 $z_{1-\alpha/2}$ 之值, 因此信賴區間便可決定了。當 α =0.1,0.5 及0.01 時(這是幾個常取的 α 值), $z_{1-\alpha/2}$ 之值分別約為1.64,1.96 及2.576。

例1.某工廠生產某種花瓶, 根據過去的經驗, 瓶口直徑(單位為公分)有常態分佈, 標準差為1.1。從某日的產品隨機抽取10個, 量其直徑分別為7.2,8,7.3,6.9,7.3,7.0,7.1,7.5,7.1,7.8。試給出直徑之期望值的一 95% 信賴區間。

解.在此 $X_1, X_2, \cdots, X_{10}$ 為 i.i.d. 之 $\mathcal{N}(\mu, 1.1^2)$ 隨機變數, μ 即為其期望值。因 $\overline{X}_n = 7.32$,$\alpha = 0.05$, 故 μ 之一 95% 信賴區間為

$$[7.32 - 1.1 \times 1.96/\sqrt{10}\ ,\ 7.32 + 1.1 \times 1.96/\sqrt{10}],$$

約為[6.64 , 8.00]。

由以上的討論知, 當母體有常態分佈且標準差已知時, 要求出期望值 μ 之一信賴區間, 可說是一極簡易的問題。當然你可能會問, 在很多實例中, 母體之標準差 σ 可能為未知, 此時該如何?在統計學裡, 樣本變異數(sample variance)

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

常用來做為 σ^2 之估計值。當樣本數夠大, 則可以 S_n 取代(4)式中之 σ, 而得近似之信賴區間。樣本數有多少才算夠大呢?通常超過30就可以了。又若 σ 未知而樣本數又不超過30該如何呢? 統計學裡也指出, 利用 $(\overline{X}_n-\mu)/(S_n/\sqrt{n})$ 有自由度 (degree of freedom) $n-1$ 的 t 分佈, 以 $\mathcal{T}_{n-1}$ 表之, 再利用假設檢定裡的一些結果仍可得到信賴區間, 這也牽涉到非常態分佈的信賴區間該如何求出, 細節在此不多討論(假設檢定與信賴區間二者間, 有一對應關係。一般而言每一信賴區間便對應一檢定, 反之亦然, 在第4節裡我們會提到)。

我們既然不擬對信賴區間多做深入探討, 那究竟要討論什麼呢?

首先所謂 $100(1-\alpha)\%$ 信賴區間到底是什麼意思? 當你看到有人指著一區間說, 這是 μ 的 95% 信賴區間, 他是在信賴什麼呢? 95% 的意義又是什麼呢? 對於(4)式, 我們通常說有 $100(1-\alpha)\%$ 的信心 μ 會屬於區間 I。但對於例1, 我們是否可說 μ 會落在區間[6.64,8.00]之機率約為0.95? 不少人以為此答案是肯定的。事實上, 對於例1, 敘述 $P(\mu\in[6.64,8.00])\doteq 0.95$ 並不正確。(4)式為一隨機區間, 在取樣前, 有 $1-\alpha$ 的機率, 此區間會包含 μ。但是一旦取得一組樣本 $x_1,x_2,\cdots,x_n$, 且將(4)式中之 $\overline{X}_n$ 以 $\bar{x}_n=\sum_{i=1}^n x_i/n$ 取代, 則所有隨機性便消失了, 而是得到一特別的區間。又因 μ 為一常數(只是不知其值為何), μ 要嘛落在此區間, 要嘛不落在此區間, 說 $P(\mu\in[\ \bar{x}_n-\sigma z_{1-\alpha/2}/\sqrt{n}\ ,\ \bar{x}_n+\sigma z_{1-\alpha/2}/\sqrt{n}\ ])=1-\alpha$ 自然不對。例如, 在例1中, 若該工廠一資深員工知道 μ 應很接近8.1, 則若你告訴他 $P(\mu\in[6.64,8.00])=0.95$, 他一定斥為無稽(此正如設一袋中有1個紅球9個白球, 某人隨機地取一球, 設取中紅球。這時你告訴他此球為白球之機率為0.9, 他必覺得你不知所云)。但在同一 α 及 n 之下, 若我們持續地取樣, 每次各得一信賴區間, 則長期而言, 這些信賴區間中, 約有 $100(1-\alpha)\%$ 個會涵蓋 μ 值。

藉圖形來說明。圖1 為 $\mathcal{N}(\mu,\sigma^2)$ 分佈之樣本平均$\overline{X}_n$之p.d.f.的圖形, 圖2 為在 σ 已知之下, 依序取樣14 次(每次皆取 n 個樣本), 所得之14個 95% 信賴區間。若 $\overline{X}_n$ 介於 $\mu-1.96\sigma/\sqrt{n}$ 及 $\mu+1.96\sigma/\sqrt{n}$ 之間, 則得到的信賴區間會包含 μ。由於圖1中機率密度函數的圖形介於 $\mu-1.96\sigma/\sqrt{n}$ 與 $\mu+1.96\sigma/\sqrt{n}$ 間之面積約為0.95, $\overline{X}_n$ 會落在此範圍內之機率便也約為0.95。圖2中之14個信賴區間, 第9 個並未包含 μ值。但由頻率對機率的解釋(frequentistic interpretation of probability), 我們知道若取樣夠多次(因此得到很多 $\overline{X}_n$ 之實際值 $\bar{x}_n$), 則其中約有 95% 個(口語有時講二十次中有十九次)信賴區間會包含 μ。至於對任一特別的區間, 說其有 95% 的機率會包含 μ, 則是沒有意義的。因這一特定的區間已非隨機區間, 常數(但未知)μ 會落在此區間的機率不是1便是0。

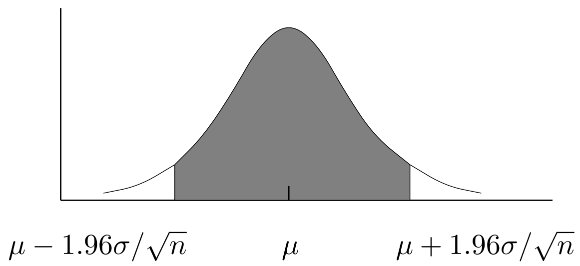

圖1 $\overline{X}_n$之機率密度函數的圖形

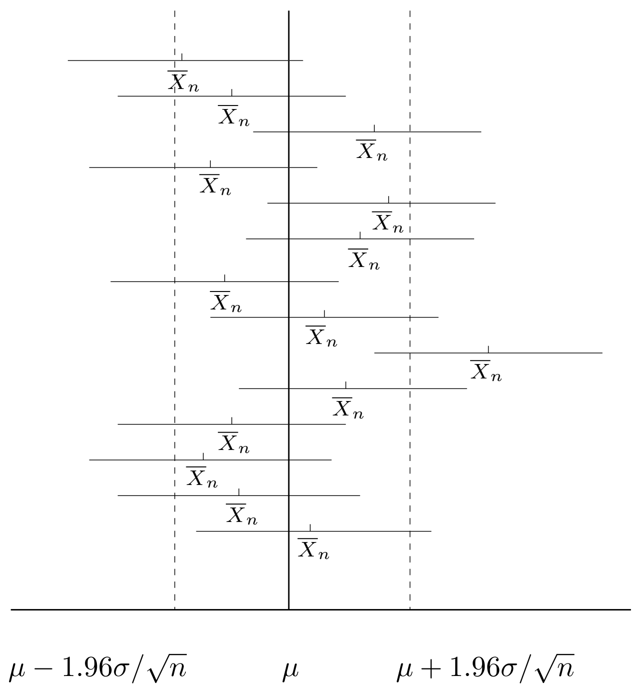

圖2 對 $\mathcal{N}(\mu, \sigma^2/n)$ 經由重複取樣所得之14個 95% 信賴區間

3. 信賴區間與品質管制

仍以常態分佈為例 。在上節中, 我們說(4)式給出 μ 之"一"信賴係數為 $1-\alpha$ 之信賴區間, 此即隱含對同一 α 值, 信賴區間並不唯一 。例如,

$$I_1 = \left[\ \overline{X}_n - \sigma z_{1-2\alpha/3}/\sqrt{n}\ ,\ \overline{X}_n + \sigma z_{1-\alpha/3}/\sqrt{n}\ \right]$$

亦為 μ 之一信賴係數為 $1-\alpha$ 之信賴區間 。只是由圖1可看出(假如你有些幾何的概念), 區間 I_1 的長度大於(4)式中區間 I 之長度 。一般而言, 信賴區間的長度愈短愈好, 愈短表愈精確 。假設你給出一參數 μ 之 95% 信賴區間, 若此區間極長, 說不定有人會取笑你乾脆取為 $(-\infty, \infty)$, 保證 μ 落在此區間 。

有些工廠的品質管制(以下簡稱品管)人員, 便以(4)式中之信賴區間, 作為品管之依據 。假設某產品之規格(如長度 、重量等)須為 μ 。經隨機抽取 n 個樣本後, 得到一如(4)式之 95% 信

賴區間(σ 假設為已知)。則若 μ 落在此區間, 便認為該批產品為合格, 否則認為不合格。偶而有品管人員心存疑惑, 取樣愈多(n 愈大), 則信賴區間的長度愈短((4)式中之區間長度為 $2\sigma z_{1-\alpha/2}/\sqrt{n}$,隨著 n 增大而變小), 因此愈不容易包含 μ , 如此一來, 不是產品愈容易不合規格嗎? 所以他們對取樣愈多存有抗拒之心(更何況取樣多本來就已較麻煩)。你認為他們的抗拒合理嗎? 檢視圖2, 區間長度若較短, 則會包含實際 μ 值之機會的確是較小, 品管人員之排斥較大的 n 似乎是有道理的。

你若同意那些品管人員的看法, 那眞是印證若要將統計學好, 還是得先對機率下些功夫, 否則不過學些花拳繡腿, 經常停留在見山不是山, 見水不是水的階段。

若採用(4)式做為信賴區間, 對一固定產品(因此 σ 相同), 在同一 α 值之下, 信賴區間隨著 n 之增大而變短。但不要忘記, α 沒有改變, 換句話說, 對這些或長或短的區間, 我們皆有相同的 $100(1-\alpha)\%$ 信心, 認為 μ 會落在其中。所以就理論而言, 認為 n 較大時, μ 便較不易落在對應之信賴區間, 其實是沒有道理的, 故此實為不成問題的問題。不若為何1加1等於2?或為何複數不能比大小? 都還成為一問題。但我們還是願稍加解釋, 免得你嘴裡不説, 心裡卻懷疑統計是否有乾坤大挪移之功。

由大數法則(Law of large numbers) 知, n 愈大時, $\overline{X}_n$ 有愈靠近 μ 之傾向(這是白話, 正式的説法請查一般機率論的書), 即“在某種意義下”,隨著 n 之變大, $\overline{X}_n$ 會趨近至 μ。因此 n 愈大時, 以 $\overline{X}_n$ 為中心, 只需要較小的半徑(長度為 $\sigma z_{1-\alpha/2}/\sqrt{n}$), 該區間仍有相同的機率 $1-\alpha$ 會涵蓋 μ。有點像若飛彈射得愈準, 則雖爆破半徑較小, 對目標物仍可有相同的摧毀效果。

其次就是涵蓋 μ 的機率若相同,我們仍較偏好區間長度較小者。原因很簡單, 區間長度愈短, 表示推論愈精準。你告訴別人 μ 有0.95的機率會落在一長度為 10 的區間, 也同樣有 0.95 的機率會落在另一長度為 5 的區間, 一般人當然覺得後者較準。這就是取樣較多(n 較大)所換得之代價。下例亦顯示信賴區間太大之缺失。

例2.在某項選舉中有兩位候選人, 欲了解選民對其中某一候選人之支持程度。隨機抽樣50人, 發現其中有27人支持該候選人。問此時該候選人是否可安心地以為穩操勝券?

解.假設選民總數夠多, 因而可忽略取樣後不放回(sampling without replacement), 與取樣後放回 (sampling with replacement)間之差異。則本問題可採用下述機率模型:設 $X_1, X_2, \cdots, X_n$ 為i.i.d.以 $\mathcal{B}(1,p)$為其共同分佈。我們擬給一 p 之 95% 信賴區間。利用中央極限定理(在一般的實例中, 通常樣本數30以上, 以中央極限定理來估計誤差便不大了), 得 $(\overline{X}_n-p)/(\sigma/\sqrt{n})$ 有近似的 $\mathcal{N}(0,1)$ 分佈。其中 $\sigma=\sqrt{p(1-p)}$ 為 $\mathcal{B}(1,p)$ 分佈之標準差, 此處當然是未知。利用 $n\to\infty$ 時,$\overline{X}_n(1-\overline{X}_n)$ 機率收斂(converge in probability)至 $p(1-p)$, 便得 p 之一近似

的 $100(1-\alpha)\%$ 信賴區間為

$$[\ \overline{X}_n - \sqrt{\overline{X}_n(1-\overline{X}_n)/n}\ z_{1-\alpha/2}\ ,\ \overline{X}_n + \sqrt{\overline{X}_n(1-\overline{X}_n)/n}\ z_{1-\alpha/2}]\ 。$$

因 $\bar{x}_n = 27/50 = 0.54$, 且 $\sqrt{\bar{x}_n(1-\bar{x}_n)} = \sqrt{0.54 \times 0.46} \doteq 0.498$, 故 p 之近似的 95% 信賴區間為

$$\begin{aligned}&[0.54 - 0.498 \cdot 1.96/\sqrt{50}, 0.54 + 0.498 \cdot 1.96/\sqrt{50}]\\ &\doteq [0.402, 0.678]\ 。\end{aligned}$$

由於上述區間包含小於0.5 的部分[0.402,0.5), 且長度並不算短, 故雖事先的抽樣顯示該候選人的支持度較高, 在選舉時若該候選人落敗並不足為奇 。

但若取樣增加為 n =1,000, 且得到540 個支持者, 則 $\bar{x}_n$ 仍為0.54, 但此時 p 之近似的 95% 信賴區間成為[0.510,0.571], 區間長度不但變短, 且0.5落在此區間之左側 。因此在相等的 $\bar{x}_n$, 且同樣的信賴係數之下, 後者(n 較大,區間較短)顯然給我們一個較精確的推論 。

4. 從信賴區間至假設檢定

我們先看下例 。

例3. 假設某電池壽命(單位為小時)有常態分佈, 標準差 $\sigma = 2$,電池壽命之期望值 μ 要等於100才符合要求 。今對某批產品抽檢 10 個樣本, 分別測量其壽命為101,103,107,99,102,101, 96,104,99,103 。試問 μ 是否為 100 小時?

解.依假設即知 $X_1, X_2, \cdots, X_n$ 為i.i.d.之 $\mathcal{N}(\mu, 2^2)$ 隨機變數, 而想回答 $\mu = 100$ 是否正確 。如果 $\mu = 100$ 之假設成立, 則 X_i 有 $N(100, 2^2)$ 分佈, $i = 1, 2, \cdots, n$ 。現考慮統計量

$$Z_n = \frac{\overline{X}_n - 100}{2/\sqrt{n}}\ 。$$

則 Z_n 有 $\mathcal{N}(0,1)$ 分佈 。因此對 $\forall 0 < \alpha < 1$,

$$P\left(\left|\frac{\overline{X}_n - 100}{2/\sqrt{n}}\right| > z_{1-\alpha/2}\right) = \alpha\ 。$$

若取 $\alpha = 0.05$, 則 $z_{1-0.05/2} = z_{0.975} \doteq 1.96$, 且取樣之 $\bar{x}_{10} = 101.5$, 而 $(\bar{x}_{10} - 100)/(2/\sqrt{10}) \doteq 2.37 > 1.96$ 。也就是小機率事件

$$\left|\frac{\overline{X}_{10} - 100}{2/\sqrt{10}}\right| > z_{0.975}$$

(此事件之機率僅約為 0.05)竟然發生了。因此我們可合理地推測原假設 $\mu = 100$ 不成立。又前述 Z_n 稱為此檢定之檢定統計量(test statistic)。

如果在另一次取樣得到 $\bar{x}_{10} = 101.1$, 因

$$\left|\frac{101.1 - 100}{2/\sqrt{10}}\right| \doteq 1.73 < z_{0.975},$$

此時我們便不否定原假設 $\mu = 100$。

在此我們給一些假設檢定之基本概念ĉ我們將例3中的假設 $\mu = 100$ 記作

$$H_0 : \mu = 100,$$

並稱此為虛無假設(null hypothesis), 而把 $\mu \neq 100$ 稱作對立假設(alternative hypothesis), 記作

$$H_a : \mu \neq 100。$$

視情況之不同, 可有不同的虛無假設及對立假設。如 $H_0 : \mu \leq 100$,且$H_a : \mu > 100$等。例3中的α 稱為顯著水準(level of significance, 或significance level)。拒絕虛無假設的區域稱為拒絕域(rejection region 或critical region), 或棄卻域。若一觀測值落在拒絕域中, 則稱該觀測值於水準 α 之下, 有統計顯著性(statistically significant at level α)。當 $\alpha = 0.05$,例3中之拒絕域為

$$\{|(\overline{X}_n - 100)/(2/\sqrt{n})| > z_{0.975}\},$$

即

$$(-\infty, 100 - 2z_{0.975}/\sqrt{n}) \cup (100 + 2z_{0.975}/\sqrt{n}, \infty),$$

當 $\overline{X}_n$屬於此區域便拒絕H_0。拒絕域之餘集則稱接受域(acceptance region), 即 $\{|(\overline{X}_n - 100)/(2/\sqrt{n})| \leq z_{0.975}\}$。

對於例3, 你可能會說拒絕或接受 H_0, 似乎有些是依"運氣"。的確如此, 更明確地說, 我們可能會犯下述兩種錯誤之一:在 H_0 為眞之下拒絕 H_0, 或在 H_a 為眞之下接受 H_0。前者稱為第一型錯誤(type I error), 後者稱為第二型錯誤 (type II error)。分別以 α, β 表犯第一型及第二型錯誤的機率。我們當然希望犯此兩種錯誤的機率同時都很小。但一般而言, 在樣本數 n 固定之下, α 愈小則 β 愈大, 反之 β 愈小則 α 愈大, 並無法讓 α 及 β 同時都變小。例如, 若有一檢定, 其 α 為 0, 此表永遠接受 H_0 ,因此 $\beta = 1$。在實際應用時, 通常我們先控制 α 值(底下會説明為何先控制 α 值), 且在給定一 α 之下, 找一 β 值最小的檢定法,也就是給出拒絕域。

這其間的細節我們不討論了, 各位可參考一般統計學的書。我們僅提出幾個須留意的要點。

首先雖然我們採用"接受"及"拒絕"的字眼, 但必須了解的是, 拒絕一假設 H_0 表認為 H_0 不成立, 然而接受 H_0 往往僅表沒有充分的證據顯示 H_0 不成立。通常在做檢定時, 要將"想拒絕之看法"置於 H_0, 或者說將"想接受之看法"置於 H_a。例如, 某廠牌之輪胎平均可行駛 $30,000$ 公里, 現製出一新型輪胎, 想判定是否較舊型為優。令 μ 表新型輪胎行駛里程之期望值, 則可取 $H_0 : \mu \leq 30,000$, 即假設 μ 並未提高, 且取 $H_a : \mu > 30,000$。因我們是想推翻 H_0, 即採認新型輪胎較優。各位應也可明白為何 H_0 稱為虛無假設了。因接受 H_0, 即認為新型輪胎並不優於舊輪胎, 這種結論實在是沒什麼好公佈的(一般人是不會公佈失敗的結果), 主持這項檢定者, 對接受 H_0 是毫無喜悅可言。

從以上的說明, 可看出假設檢定乃採用反證法。即先假設某不想要的情況成立。然後在此假設下進行推導, 如果得到矛盾, 則便推翻原來的假設。如果沒有得到矛盾, 則便不拒絕原來的假設。但是此處的反證法, 有別於數學中的反證法。因此處所謂矛盾, 並非形式邏輯中的絕對矛盾, 而是基於人們在實際經驗中常採用的一原則: 小機率事件, 在一次實驗中不易發生。根據此一原則, 如果小機率事件在一次實驗中發生了, 便認為原來的假設不成立。換句話說, 假設檢定中所採用者, 乃是一種機率式的反證法。

底下給一例來說明為何 α 與 β 無法同時變小。

例4. 設某藥有0.25的機率能治癒某種疾病, 現有一種新的且較貴的藥, 想檢定此新藥是否優於舊藥。將新藥讓20個得此病的病人服用, 令 X 表治癒的病人數, 則 X 有 $\mathcal{B}(20, p)$ 分佈, 其中 p 表治癒率。由於目的是希望能宣稱新藥較優(新的藥且又較貴, 如果不是較優又何需生產), 所以將$p = 0.25$置於H_0,H_a則為$p > 0.25$, 且若有較多的病人被治癒, 則認為 H_a 為真。即我們要檢定

$$
\begin{aligned}
H_0 &: p = 0.25, \text{且} \\
H_a &: p > 0.25,
\end{aligned}
$$

而拒絕域取為$\{X \geq c\}$。當 $c = 9$,則第一型錯誤的機率為

$$
\begin{aligned}
\alpha &= P(X \geq 9 | p = 0.25) \\
&= 1 - \sum_{i=0}^{8} \binom{20}{i} (0.25)^i (0.75)^{20-i} \\
&\doteq 1 - 0.9591 = 0.0409,
\end{aligned}
$$

為一個很小的值。譬如說實際觀測到$X = 10$, 則接受新藥較優的看法。

至於第二型錯誤的機率β, 對此模式並無法計算出來, 除非在H_a中, p 為一個明確的值 。現若改為$H_a: p = 0.5$, 則

$$\beta = P(X < 9|p = 0.5) = \sum_{i=0}^{8} \binom{20}{i} (0.5)^{20} \doteq 0.2517 \text{ 。}$$

此機率並不算小, 此結果顯示在新藥明顯優於舊藥之下(治療率為二倍), 卻有超過四分之一的機率, 會拒絕新藥較佳 。

在很多統計軟體裡, 於執行一假設檢定時, 會算出所謂 p 值(p-value) 。
然後看你所選的 α 值為何, 若 $p < \alpha$ 則拒絕 H_0 。對一觀測值, 所謂 p 值, 乃在 H_0 為眞之下, 檢定統計量會等於該觀測值, 或較該觀測值更極端的機率(也就是會使該觀測值導至拒絕 H_0 之最小的 α) 。如前, 若觀測到 $X = 9$, 則 p 值為0.0409; 若觀測到 $X = 10$, 則 p 值為0.01385 。可看出若 p 值愈小, 則觀測值所提供拒絕 H_0 的證據就愈強, 換句話説愈顯著 。

若H_0仍不變, H_a改為$p = 0.7$會如何? 亦即除非新藥之優勢更高, 否則寧採舊藥 。此時

$$\beta = P(X < 9|p = 0.7) = \sum_{i=0}^{8} \binom{20}{i} (0.7)^i (0.3)^{20-i} \doteq 0.0051,$$

降低很多 。

最後我們再看, 在$H_0: p = 0.25$,且$H_a: p = 0.5$之下, 若拒絕域改為$\{X \geq 8\}$, α,β會有什麼改變? 此時

$$\alpha = 1 - \sum_{i=0}^{7} \binom{20}{i} (0.25)^i (0.75)^{20-i} \doteq 1 - 0.8982 = 0.1018,$$

而

$$\beta = \sum_{i=0}^{7} \binom{20}{i} 0.5^{20} \doteq 0.1316 \text{ 。}$$

採用此新策略,β變小了, 但α卻變大了 。不難看出若 c 增大, 則 α 變小且 β 變大; 反之若 c 變小, 則 α 變大且 β 變小 。特別地, 當 $\alpha = 0$(即 $c > 20$)時, $\beta = 1$; 當 $\beta = 0$ 時(即 $c < 0$),$\alpha = 1$ 。印證如前所述, 在樣本數固定之下, 對同一個H_0及H_a, 通常減少某一型之錯誤, 便增大另一型錯誤 。

再看一個類似的例子 。

例5. 一般認為生男的機率為 $p = 0.5$, 不過數據顯示 $p > 0.5$(如對高加索人(Caucasian)p 約為0.512) 。欲證實 $p > 0.5$, 隨機地取$n = 10$個初生嬰兒 。則其之中男孩數X有$\mathcal{B}(10, p)$分佈 。要檢定$H_0: p = 0.5$,且$H_a: p > 0.5$ 。如上例當男孩數較多時拒絕H_0, 因此拒絕域為$\{X \geq c\}$, 其中 c 要選的夠大, 使得第一型錯誤的機率α 夠小 。表一為$p = 0.5$時, X 之機率密度函數, 其中 $p(i) = P(X = i|p = 0.5)$ 。

對每一 c, 由表1可求出 $\alpha{=}P(X\geq c|p=0.5){=}p(c)+p(c+1)+\cdots+p(10)$之值。當 c 分別等於5,6,7,8,9,10 時,α分別約為0.62306、0.37697 、0.17189、0.05470、0.01075、0.00098。若取 $c=8$,則 $\alpha\doteq 0.05470$, 即若拒絕域為 $\{X\geq 8\}$, 則 $\alpha\doteq 0.05470$, 並不算大。看到這裡, 大部分的人會覺得此檢定的過程還算合理。另一方面, 若H_a為眞, 且$p=0.512$, 則β為

$$\begin{aligned}&P(X<8|p=0.512)\\&=1-\binom{10}{8}(0.512)^8(0.488)^2-\binom{10}{9}(0.512)^9(0.488)-\binom{10}{10}(0.512)^{10}\\&\doteq 0.9364\text{。}\end{aligned}$$

第二型錯誤的機率將近1。換句話說, 在H_0不眞之下, 我們仍有極大的機率接受H_0。即不論H_0 為眞或不眞, 均極易接受H_0為眞, 因此這個檢定過程顯然並不合理。

表1 $\mathcal{B}(10,0.5)$之機率密度函數

i	0	1	2	3	4	5
$p(i)$	.00098	.00977	.04395	.11719	.20508	.24609

i	6	7	8	9	10
$p(i)$	.20508	.11719	.04395	.00977	.00098

事實上, 若欲使α約為0.05, 且β不超過0.1, 則要有更多的數據(即 n 要較大才行)。你要不要猜究竟 n 要多大? 可能要嚇你一跳, n 大約要15,000以上才行。主要是0.5與0.512太接近之故。

本例告訴我們千萬不要只看到α不大, 就貿然進行一檢定, 需也檢視β之值的大小。

在 $H_0:\mu=\mu_0$, 且 $H_a:\mu\neq\mu_0$之下, 利用信賴區間可得一檢定法。設母體有$\mathcal{N}(\mu,\sigma^2)$分佈, σ已知, 則可取接受域為如(4)之信賴區間。換句話說, 若$\mu\in I$ 則接受H_0, 若$\mu\not\in I$ 則拒絕H_0。舉例來說明, 設某工廠對某產品的要求為$\mu=7.0$。該工廠品管人員的作法是取樣後, 若$7.0\in I$, 則認為該批產品合格, 反之則認為不合格, 你認為他們的作法是否正確?

其實是不對的。用白話講放在H_0的敘述是要被保護的, 沒有充分的證據不輕易推翻, 接受H_0往往是無可奈何, 因證據不足, 並不是眞相信H_0就一定是最好的選擇。對於諸如消費者文教基金會, 若他們懷疑前述工廠的產品有問題, 想做一個檢定, 則沒錯, 要取 $H_0:\mu=7.0$,且 $H_a:\mu\neq 7.0$。寧可先相信該工廠, 則一旦抽樣檢定拒絕 H_0, 該工廠就很難抗辯了。但該工廠若做品管, 也取同樣的 H_0 及 H_a, 則在接受H_0時如何取信大眾呢? 因為 H_0 是極容易被接受的。那麼該工廠要如何選擇 H_0 及 H_a 呢? 當然是

取 $H_0: \mu \neq \mu_0$,且 $H_a: \mu = \mu_0$。則當拒絕H_0, 接受H_a時, 自然可信心十足的宣稱該產品符合規格。雖然工廠及消費者文教基金會的目的都是想知道究竟$\mu = 7.0$是否為眞, 但所設的H_0及H_a卻恰好相反。簡言之, 要將希望得到的結論之反面置於H_0。此中原委是在進行一統計檢定前不可不留意的。對消費者文教基金會而言, 如果明明 H_0 是對的(產品合格), 卻被拒絕(認為產品不合格), 這當然是會引起很大爭端, 所以這種錯誤的發生要越少越好(即 α 值要控制得小些。至於 α 要多小當然也是視情況而定。如在 $\alpha = 0.05$ 之下, 指控產品不合格, 工廠有時是不太服氣的, 此時便宜取較小的 α 值)。但若 H_a 是對的(產品不合格), 卻接受 H_0(產品合格), 工廠雖一時僥倖被放過, 但夜路走多後, 難免會遇到鬼, 總有逮到該工廠產品不合格的一天(這是為何有時 β 值雖很大, 我們仍可容忍的原因)。

附帶一提, 在法律上秉持"被告在被證明有罪之前皆為清白"之原則。很多事實上有罪之被告, 便因證據不夠充分而被開釋。政治人物被法庭宣判無罪開釋時, 往往很高興地說"司法還我清白"。如果了解法庭其實是取H_0: 無罪, 且H_a :有罪, 便不會把"無罪的宣判"與"眞正無罪"劃上等號了(再回頭看一下例4及例5, β值(在此即有罪卻誤判無罪之機率)有時會很大的)。

宋朝歐陽修在追述其父母生前言行事蹟的瀧岡阡表一文(收錄於古文觀止), 提及其父治死獄的情形"求其生而不得, 則死者與我皆無恨也。"也是這種先相信對方(工廠、被告等)的精神。歐陽修又寫著"夫常求其生, 猶失之死, 而世常求其死也。"更是值得我們警惕。保持開放的態度, 不要有先入為主的偏見, 不論在法庭上、在假設檢定裡, 甚至在整個人生中, 均是適用的。

此道理一經說出, 彷如老生常談, 實際上卻並不易做到。要知自以為是, 及自以為較別人行, 為一般人的通病。上至國民大會的多次修改憲法, 下至有些單位換新主管後, 便將原有的制度(或各種辦法)修改, 甚至整個推翻。事實上, 若能虛心些, 謹守尊重現況(即置現況為 H_0, 不輕易否定)的精神, 則於制定(或修改)一項辦法前, 大家會更愼重, 沒有充分的把握寧可不改變現況。並且因知一旦制定(或修改)後, 將來又是不易被修改, 會使大家下決定前, 能更考慮周全。古人批評"朝令夕改", 今人說"朝令有錯, 夕改何妨?"想想古人還眞有智慧。

5. 結語

本文只是就信賴區間及假設檢定, 討論一些一般人容易誤用之處。統計推論為統計學中一重要的題材, 學習統計後對那些琳瑯滿目的方法, 使用前要先了解其中的含意, 才能眞正發揮統計的功能。至於完整的信賴區間及統計檢定的討論, 可參考諸如Roussas(1997)等統計學的書籍。

習　題

1. 設 X 有 $\mathcal{N}(0,\sigma^2)$ 分佈, $\sigma > 0$。求隨機區間 $(|X|, 10|X|)$ 包含 σ 之機率, 並求此區間長度之期望值。

2. 設 $X_1, X_2, \cdots, X_n$ 為i.i.d.之隨機變數, 以 $\mathcal{U}(0,\theta)$ 為其共同分佈, 又令 $Y = \max\{X_1, X_2, \cdots, X_n\}$。給二常數 a, b, $1 \le a < b$, 以 $[aY, bY]$ 做為 θ 之區間估計。求此區間之信賴係數。

3. 設 $X_1, X_2, \cdots, X_n$ 為i.i.d.之隨機變數, 以 $\mathcal{N}(\mu,\sigma^2)$ 為其共同分佈, μ,σ皆為未知。經取樣得 $n = 10$,$\bar{x}_{10} = 3.22$,$s_{10} = 1.17$。求 μ 之一 95% 信賴區間。

4. 設 $X_1, X_2, \cdots, X_n$ 為i.i.d.之隨機變數, 以 $\mathcal{P}(\lambda)$ 為其共同分佈。試以 $\overline{X}_n$ 給出一 λ 之近似的 $100(1-\alpha)\%$ 信賴區間。

5. 對於例5, 當 n =15,000時, 利用二項分佈趨近至常態分佈, 求 c 值使得 α 約為0.05, 並求此時之 β 值。

6. 設 X 有 $\mathcal{B}(5,\theta)$ 分佈, $0 \le \theta \le 1$。欲檢定 $H_0 : \theta \le 1/2$, 且$H_a : \theta > 1/2$。取拒絕域為 $\{X > i\}$,$i = 0, 1, \cdots, 5$, 且令 $\beta_i(\theta) = P_\theta(X > i)$ (對每一 i,$\beta_i(\theta)$ 稱為以 $\{X > i\}$ 為拒絕域之檢定的強力函數(power function))。
   (i)對 $\forall\, i = 0, 1, \cdots, 5$, 寫出 $\beta_i(\theta)$。
   (ii)對一固定的 i, 說明 $\theta \le 1/2$ 時, $\beta_i(\theta)$ 表第一型錯誤, 而 $\theta > 1/2$ 時,$1 - \beta_i(\theta)$表第二型錯誤。
   (iii)試藉 $\beta_4(\theta) \le (1/2)^5$, $\forall\, \theta \le 1/2$, 且只有當 $\theta > (1/2)^{1/5}$ 時,$\beta_4(\theta) > 1/2$, 說明當拒絕域為 $\{X > 4\}$ 時, 第一型錯誤皆很小; 而對大部分的 $\theta > 1/2$, 第二型錯誤皆不小。
   (iv)試繪 $\beta_2(\theta), \beta_3(\theta)$ 及 $\beta_4(\theta)$ 之圖形, 並分別說明對 i=2,3,4,第一型及第二型錯誤之增減情況。

參考文獻

1. Roussas, G.G.(1997). *A Course in Mathematical Statistics*, 2nd.ed. Academic Press, San Diego.