統計顯著性

黃文璋

1. 尋常

朱雀橋邊野草花, 烏衣巷口夕陽斜;
舊時王謝堂前燕, 飛入尋常百姓家。

這是唐詩三百首裡, 劉禹錫的烏衣巷。烏衣巷是東晉王導、謝安兩望族居住的地方, 當時很繁榮, 曾幾何時沒落了, 只見野草花, 夕陽斜。從前王、謝兩家堂前的燕子, 如今已飛入平常百姓的家。暗指王、謝豪門的子弟, 已淪爲尋常百姓。

尋常就是平常。而非比尋常, 不同尋常, 或簡單地說不尋常, 當然是尋常的反義字。一件尋常的事, 就是常可見到, 常會發生, 它發生並不會讓人驚訝。不尋常的事發生, 則會令人驚訝。

只是尋常與不尋常如何區隔呢? 怎樣才是尋常百姓家? 在南方範的小說桃花扇裡, 寫明末名士侯朝宗, 與秦淮名妓李香君間, 一段淒美的愛情故事。書中李香君曾講一句"爛船還有三千釘, 畢竟是尚書府裏闊哥兒, 逃難下來, 仍捧得出三百兩白花花銀子。"落魄公子, 在別人眼裡, 不見得都是苦哈哈。所以何謂尋常, 是一相對的看法。住在昔日王、謝豪宅中的新主人, 甚至同一巷子中的居民, 想必都不會是太尋常的老百姓。

2. 顯著

有些事看起來稀鬆平常, 尋常的一件事。有些事看起來不太尋常, 在做比較時, 我們有一名詞來描述–顯著:

顯著進步, 差異顯著, 成果顯著, $\cdots$。

當然如同尋常, 怎樣才是顯著, 標準可有很大的差異。在洪蘭譯 (2005) 的頁296-297, 談到預測的誤差。如全球暖化的預測錯了3倍; 3小時的飛行航程 1 小時就到了。書中認爲這樣的差異是大的。另外, 也舉了一誤差小的例子。

> 美國太空總署發射載有登陸火星的探測船精神號時, 他們宣稱在253天後, 探測船會在加州時間下午8點11分登陸火星, 結果它在8點35分登陸, 這個誤差只有幾千分之一。美國太空總署的人知道他們在說些什麼。

實際值是預測值的3倍或1/3倍, 與誤差是幾千分之一相比, 當然是很大的。有趣的是, 登陸誤差24分, 爲一天的60分之一 (一天有1,440分), 所以誤差約是 (因不曉得幾時幾分發射, 所以只能說約是)

$$\frac{1}{60} \div 253 = \frac{1}{15,180}。$$

誤差並非如書上所說的幾千分之一, 而是15,180分之一。看來要犯 "顯著的" 誤差, 是很容易的。

怎樣的誤差算大, 怎樣算小? 怎樣的誤差是很尋常, 怎樣是很顯著? 這顯然是很主觀的。第4節我們會說明如何認定顯著。

3. 證實

數學裡常在證明, 不少人甚至以爲學數學的目的就是在學證明。所謂證明, 就是給某條件 (條件 A), 要導致某結論 (結論 B) 成立。對一如下的敍述 (亦稱爲一命題)

若 A 則 B,

此命題是否爲眞? 假設 A 成立, 若能導致 B 亦成立, 此命題便爲眞; 若無法導致 B 成立, 此命題便不眞。A 是假設成立的, 不用去證明。例如, 當你看到 "設$x > 3$", 不用去證明 $x > 3$, 而是要看 $x > 3$ 之後怎麼樣。曾有人說如果假設給的夠, 任何結論皆可成立。例如, 有一命題:

$$\text{若 } 4 = 5, \text{ 則 } 1 > 2。 \tag{1}$$

看起來很荒謬, 4 怎麼會等於 5, 1 怎麼會大於 2 ? 由於 "若 A 則 B" 等價於"若非 B 則非 A", 故 (1) 等價於

$$\text{若 } 1 \le 2, \text{ 則 } 4 \ne 5。$$

因命題 (2) 爲眞 (不論1與2之關係爲何, $4 \ne 5$永遠是對的), 故命題 (1) 亦爲眞。這類命題很多, 有人宣稱:

天塌下來我都不怕。

他很勇敢嗎? 非也, 天是不會塌下來的 (如同4不會等於5), 所以他怎麼宣稱, 該命題都是對的。

除了前述這些邏輯上的命題判斷眞僞外, 數學中不乏規規矩矩的證明。例如, 設 $a, b \ge 0$, 則

$$\frac{a+b}{2} \ge \sqrt{ab}。 \tag{2}$$

這就是著名的 *算術平均大於或等於幾何平均*, 對任二非負實數都成立的一個不等式。數學上的一個命題, 一旦證明是對的, 就毫無例外的永遠是對的。對(3) 中的不等式, 不可能找到兩個 $a, b \ge 0$, 使 (3) 不成立, 即

$$\frac{a+b}{2} < \sqrt{ab}。$$

除了在數學裡, 科學上的一項發現, 較少用證明, 而多半用“證實”。給幾個例子:

1. 英國科學家證實, 接觸殺蟲劑會使人罹患帕金森氏症的機率增加。
2. 人類基因序列破譯完成, 證實人與黑猩猩同源。
3. 英國研究人員, 證實針灸有治療效用。
4. 古法蘭西第一美女, 被科學家證實死於毒殺。
5. 考古證實, 老子生於河南鹿邑。
6. 研究證實, 素食影響女性生育。
7. 德國科學家證實, 綠茶可以減肥。
8. 全球研究證實, 家庭作業愈多考試分數愈低。
9. 三名中學生證實洗滌劑對男性生殖能力有損害。
10. 美國鉛球冠軍凱文·托特被證實服用禁藥。
11. 美國研究證實, 有性高潮更容易懷孕。

科學家以外, 也常用到以“證實”來宣佈事情。給幾個例子:

1. 美軍證實在阿富汗失蹤的海豹特遣隊員是兩死一獲救。
2. 老布希回憶, 毛澤東相信上帝, 證實中共無神論是假的。
3. 家書原件曝光, 證實釣魚台曾爲盛家藥材採集地。
4. 聯合國人權報告, 證實中共眞面目。
5. 小 S 親口證實懷孕。

證實是什麼意思? 證明其確實! 對上述這些新聞所提及之被證實的事件, 即使科學家言之鑿鑿, 恐怕仍有不少人半信半疑。至於各機構或個人“證實”的事件 (看過間諜遊戲那部電影嗎?), 更不乏是睜眼說瞎話。此與數學上的證明是完全不同的。政府的證實不談, 科學上所宣稱的“證實”, 常也是有一套自以爲合理的依據。只是對於隨機現象, 常就是無法如數學上有斬釘截鐵的結果。例如, 究竟綠茶可否減肥? 可能對某些人有效, 對某些人無效。因此需要找一些人來作實驗。而有多少比例的人體重降低, 才能證實綠茶可以減肥? 這可不是一輕易能回答的問題。另外, 是不是讓參與實驗者每人都喝綠茶呢? 顯然不行, 因這樣就看不出喝綠茶跟不喝綠茶, 對減肥的效果是否有差異。還有很多其他因素要考慮。如何進行一較客觀的實驗, 是要經過一番設計的。統計裡的實驗設計, 就是在討論這方面的問題。沒有好的實驗設計, 實驗結果的正確性, 是令人懷疑的。例如, 若對一群國中一年級的學生做喝綠茶是否能減肥的實驗, 結果很可能是不會。因國中小孩正值成長階段, 不管喝綠茶還是果汁, 或只喝水, 隔一段時間後, 體重大約都是增加。

有了一符合標準的實驗, 怎樣的結果, 才能“證實” 綠茶可以減肥, 或是新款燈泡比舊型燈泡用得更久? 現代統計學發展出一套判定的方法。

4. 假設檢定

不論"證明"或證實, 都非統計學裡做決策時的專有名詞。統計學裡, 有純粹學術探討的一部分, 但也有入世的一面。在入世的這一面, 其中的很多作法都可與我們的某種思維相對應。

在舊約聖經裡, 以智慧過人著名的 所羅門王, 對於兩婦人爭奪一嬰兒, 他如何判定嬰兒是誰的? 他要屬下將孩子劈成兩半, 各給兩婦人一半。有一婦人立即說「將活孩子給那婦人吧! 萬不可殺他。」另一婦人則說「這孩子也不歸我, 也不歸你, 把他劈了吧!」所羅門王由此判定小孩屬於第一個婦人的。以色列衆人聽見他這樣判斷, 就都敬畏他, 因爲見他心裡有神的智慧。

流傳於民間的包公審錢案 的故事大家應也聽過。包公藉由銅板扔進一盆水中飄起油花, 斷定誰偷了油條小販的錢。

現實生活裡, 能以所羅門王或包公的方式, 來決定眞相的機會其實不多。後人若想東施效顰, 恐怕誤判的機會還多些。

到底眞相爲何? 常是只有天曉得。口袋裡的錢沾到油, 便是偷自賣油條者嗎? 在一些情況下, 這不失爲一合理的猜測。例如, 偷了賣油條小販, 錢尙未用出, 且放身上, 而其他人的錢皆未沾有油。基於這種理由所見到的結果, 推測什麼情況下, 使此結果最容易發生的想法, 在統計學裡, 便發展出最大概似估計法。沒有其他資訊時, 是可以考慮採用此法。只是有時壞人較狡猾, 較會僞裝, 事情所呈現的表象, 不見得是眞實的情況, 最大概似法, 在此便可能不靈。

不論是東方或西方, 自古即知律法的重要。漢摩拉比法典, 即使不是世界上最古老的法典彙編, 也是最完備、最有條理的。此法典爲西元前1760年, 巴比倫第一個王朝的第六個國王漢摩拉比 (King Hammurabi, 西元前1792-1750年) 所頒佈的。這可說是巴比倫人所留下最重要的遺產。

有了法典, 還要判決。現今我國刑事訴訟法中 (第154條), 採無罪推定原則:

被告未經審判證明有罪確定前, 推定其爲無罪。

不能僅憑一些蛛絲馬跡, 抓了一嫌犯後, 又發現有多項嫌犯符合處, 便定嫌犯的罪。包公審錢案, 就是憑口袋裡的錢有油, 認定其爲偷錢者, 在今日無罪推定原則下, 自然都不可行了。

統計裡做判斷, 也是採無罪推定的精神, 並以 "假設檢定" (testing hypothesis) 的方式來進行。這是在西元1933年, 由波蘭人 奈曼 (Jerzy Neyman, 1894-1981), 及英國人 皮爾生 (Egon Pearson, 1895-1980), 給出著名的奈曼-皮爾生引理 (Neyman - Pearson lemma) 所奠定的。相對於我們最高法院, 於民國25年, 立下有罪推定原則判例的六十餘年後, 終於在民國92年1月, 公佈了前述無罪推定的第154條。此後被告原則上是無罪的, 不必證明自己無罪, 法官只要認爲被告罪證不足, 即可判無罪, 不必窮調查之途, 才能判被告無罪。統計學裡的無罪推定原則, 較我國刑事訴訟法, 早了整整70年。

其實我們平常在判定事物, 早就依假設檢定的方式。或者倒過來說, 統計學裡的假設檢定, 其實仍是源自於我們的思維。先看底下幾個敍述。

1. 如果你們兩位沒有作弊, 怎麼4題計算題, 錯得一模一樣?
2. 如果我們兩人不是有緣, 怎麼會國中同班, 高中又同班?
3. 如果此銅板是公正的, 怎麼可能投擲10次皆得正面?

沒作弊, 4題計算錯得一模一樣, 是很罕見的, 看來那兩個學生得好好跟老師解釋, 或準備被懲罰。對於第二個敍述, 假設高雄市某國中的某一班有5人畢業後進入高雄女中就讀, 又設高雄女中一年級有24班, 且新生是隨機地分班。則此5人中至少有2人高一同班的機率爲

$$1-\frac{24\cdot 23\cdot 22\cdot 21\cdot 20}{24^5}\doteq 0.359>\frac{1}{3}。$$

高二及高三都可能重新分班。所以對那些"好班"的學生, 國中時同班, 高中時又同班, 並非太怪異的事。至於高雄女中每屆新生中, 要找出彼此"有緣"的同學, 就更多了。對於第3個敍述, 發生的情況是, 我們懷疑該銅板不是公正。或者更明確地說, 我們傾向以爲該銅板出現正面的機率 $p>0.5$。但是呢, 我們先假設該銅板爲公正, 即 $p=0.5$。

投擲銅板, 直觀上出現的正面數過多, 與 $p>0.5$ 才較吻合。如果投擲 10 次得到 6 個正面, 是否能推翻 $p=0.5$, 而得到 $p>0.5$? 即使是公正的銅板, 投擲 10次, 不要以爲比較容易恰得5個正面。因機率爲

$$\binom{10}{5}(\frac{1}{2})^{10}=\frac{252}{1,024}\doteq 0.246<\frac{1}{4}。$$

而得到正面數多於 5 的機率爲

$$\binom{10}{6}(\frac{1}{2})^{10}+\binom{10}{7}(\frac{1}{2})^{10}+\binom{10}{8}(\frac{1}{2})^{10}+\binom{10}{9}(\frac{1}{2})^{10}+\binom{10}{10}(\frac{1}{2})^{10}=\frac{386}{1,024}\doteq 0.377,$$

此爲一不算小的機率。因此不會因得到正面數比5還多, 就覺得應該是$p>0.5$。那得到7個正面呢? 得到正面數至少是7的機率爲

$$\binom{10}{7}(\frac{1}{2})^{10}+\cdots+\binom{10}{10}(\frac{1}{2})^{10}=\frac{176}{1,024}\doteq 0.172>\frac{1}{6},$$

比 1/6 還大的機率, 仍常會發生。得到 8 個正面呢? 得到正面數至少是 8 的機率爲

$$\frac{56}{1,024}\doteq 0.055。$$

這機率就有點小了, 可以懷疑 $p=0.5$ 不眞。至於得到9個正面呢? 得到正面數至少是 9 的機率爲

$$\frac{11}{1,024}\doteq 0.0107,$$

懷疑$p = 0.5$的心會更強烈。如果得到10個正面呢? 此機率爲

$$\frac{1}{1,024} \doteq 0.000977 < 0.001。$$

這麼小的機率, 大部分的人會捨 $p = 0.5$, 而就 $p > 0.5$。

爲什麼出現 6 個正面, 我們不是只考慮恰出現 6 個正面之機率, 而是考慮至少出現6個正面之機率? 此因如果出現 6 個正面, 就認爲 $p > 0.5$, 則合理的選擇是出現正面數比 6 多, 都應認爲 $p > 0.5$。其次要說明的是, 即使出現10個正面, 雖此機率小於千分之一, 但是否能就百分之百肯定 $p > 0.5$ 呢? 當然不行, 凡是正的機率, 即使再小, 都有可能發生。你只能強烈地懷疑 $p = 0.5$, 或者說強烈地相信 $p > 0.5$, 但眞相究竟爲何, 很可能並無法得知。

根據以上思維, 奈曼-皮爾生提出了一套假設檢定的架構。在其架構裡, 有一*虛無假設* (null hypothesis), 常以 H_0 表之; 及一*對立假設* (alternative hypothesis), 常以 H_a 表之。虛無假設通常表現況, 而對立假設表我們傾向相信的。例如, 對喝綠茶能減肥的問題, 生產的公司當然希望答案是肯定的, 於是會將 H_0, H_a 分別取爲

$$H_0 : \text{喝綠茶不能減肥},$$
$$H_a : \text{喝綠茶能減肥}。$$

而如前所述, 經過一實驗後, 我們不會說 *得證* 喝綠茶不能 (或能) 減肥, 而會說 *接受* (或拒絕) H_0。在統計裡, 一個關於母體之敍述, 就稱爲一 *統計假設* (statistical hypothesis)。虛無假設及對立假設皆爲統計假設。對於一統計假設, 我們要去檢定是否接受或拒絕, 這整個過程, 便稱*假設檢定*, 或稱*統計檢定*, 或簡稱 *檢定*。而導致接受或拒絕一統計假設的步驟, 就是 *統計推論* (ststistical inference) 之主要工作。

虛無假設是要特別被保護的。想想若喝綠茶明明不能減肥, 卻宣佈有效; 喝咖啡明明沒事, 卻宣佈易致癌; 明明無辜, 卻判定他殺人, 後果都很難彌補。要再度提醒各位的是, 不論虛無假設, 或對立假設, 都只是假設, 而非如數學中的命題。對於命題, 我們可以證明它是眞或僞。對於假設, 就是決定接受或不接受。接受不表示就完全相信該假設爲眞。有可能是雖不滿意但可以接受, 也有可能是無可奈何地接受。生活中也常是如此。找對象時東考慮西考慮, 最後決定嫁誰, 也只是認爲可以接受而已。接受是否就表示找到眞命天子? 當然未必, 標準降低就愈容易接受。另外一點, 對於進行一假設檢定, 比較希望的其實是拒絕H_0, 接受 H_a。如果接受 H_0, 表此實驗白做, 接受一空的 (虛無) 假設。誰會宣佈喝綠茶不能減肥呢? 誰會宣佈家庭作業愈多考試分數不會愈低呢? 這類結論引不起太多人的興趣。

在一統計檢定裡, 不論接受或拒絕 H_0, 都有可能犯錯。這其中有兩型錯誤, 我們列在表1。理想狀況當然是兩型錯誤之機率 (注意我們談的是錯誤機率, 因不論任何決策, 總是有時正

確, 有時錯誤, 只能看其機率大小) 皆爲0。但通常不會有這種情況, 此點我們留至本節最後再說明。奈曼-皮爾生的作法是, 先給定第一型錯誤的機率, 爲一比較小的值, 然後設法決定 *拒絕域* (rejection region, 或稱 critical region), 即何時拒絕H_0。不落在拒絕域就落在 *接受域* (acceptance region)。第一型錯誤的機率, 常以符號 α 表之, 即

$$\alpha = P(\text{拒絕}H_0 | H_0\text{爲眞})。$$

上述 α 也稱爲檢定之 *顯著水準* (significance level), 如果觀測值落在拒絕域中, 我們可說所獲得之數據 (或說實驗結果) 具 *顯著性* (significant), 足以拒絕 H_0。

表1. 統計檢定之可能結果。

	H_0爲眞	H_0不眞
接受H_0	正確	第二型錯誤
拒絕H_0	第一型錯誤	正確

如果要檢定兩群學生的平均成績 μ_1, μ_2 是否有差異, 即要做

$$H_0 : \mu_1 = \mu_2,$$
$$H_a : \mu_1 \neq \mu_2,$$

之檢定。在某一 α 下, 若拒絕 H_0, 則可說在顯著水準 α 下, 兩群學生的平均成績, 有顯著差異。或說結果有顯著性。在不同的情況下, 有時會說顯著提昇, 顯著改變, 顯著正相關, 抗癌效果顯著。或說差異不顯著, 療效不顯著等。

機率比較大的事件 (如投擲銅板10次得到6個正面) 就是尋常的事件, 其發生不會令人大驚小怪。機率比較小的事件 (如投擲銅板10次得到8個正面), 就是 *偶然* 事件, 看到了要有些警覺心, 可能有必要做個統計檢定。檢定結果如果拒絕 H_0 (如在 $\alpha = 0.05$ 之下, 投擲10次得到9個正面), 則檢定具顯著性。

與第2節所提口語中的"顯著"相比較, 統計裡的顯著並非指絕對值的大小, 而是指發生機率的大小。發生機率小才稱顯著。我們常說少見多怪, 也是這個意思: 見到機率小的事件, 當然會另眼相看 (顯著)。機率較大的事件, 就是尋常事件, 見到了是不會覺得奇怪的。某生模擬考進步10分, 算不算進步顯著? 如果全年級有30%的人進步至少10分, 就不算顯著; 如果只有1%的人進步至少10分, 大約便可稱進步顯著了。英文裡說"he had achieved something significant", 也是指達到的地步是較稀罕的。

有人告訴曾參的母親曾參殺人, 曾母不理會, 因社會上造謠的人很多, 她認爲此爲一尋常事件。第二個人告訴曾母曾參殺人, 曾母仍不理會。難免遇到二造謠者, 今天運氣眞不好, 她視

此爲一偶然事件。當第三個人告訴曾母曾參殺人, 曾母就逃走了。因太少一天有三個人來跟你造謠, 此爲一顯著事件, 她不得不接受她的愛子殺人。

常採用的顯著水準 α 爲 0.1, 0.05, 及0.01等, 但也可採其他的值, 看犯第一型錯誤後果之嚴重性而定。α 值愈小, 就愈不容易拒絕 H_0。法庭上若遇到攸關嫌犯生死之案件 (無罪推定, H_0 取爲嫌犯無罪), α 可能要取的更小。舉一例來看。

例 1. 陪審團制度是美國審判中的一大特色。通常由12人組成的陪審團, 在審判中聽取原告、被告雙方的陳述, 證人的證詞, 查看證據, 聆聽法官對法律的解釋, 最後做出被告是否有罪的裁決。一般的民事或刑事案件要求贊成票達9票以上, 指控謀殺的案件, 則要求一致通過。

假設殺人嫌犯實際無罪, 且設任一陪審員會誤判他有罪之機率爲0.2(太高便不合理), 各陪審員之決策行爲假設爲相互獨立。則無辜者會被判決有罪之機率爲

$$0.2^{12} \doteq 4.096 \cdot 10^{-9},$$

算是很小。

要注意一點, 不論α值多小, 都可能犯第一型錯誤。以投擲銅板爲例, 若10次皆得正面, 則即使 α 小至0.001, 都會拒絕 H_0: 銅板爲公正。但我們知道, 只要實驗次數夠多 (例如找1,000個人來各投擲銅板10次), 對一公正的銅板, 觀測到10次皆得到正面, 就很尋常了。

由於顯著水準 α 該取爲多少, 並無一定準則, 視不同情況、不同人而定。而且宣佈接受H_0, 並未指出是勉強接受, 或安心地接受, 因此遂引進了 ***p*-值**(p-value) 的概念。

所謂 p-值, 乃在 H_0爲眞之下, 比觀測值至少同樣極端之區域的機率。給出 p-值後, 不同的決策者, 可依其所設定的α值, 而做出決定: p-值小於或等於 α 值便拒絕 H_0, 大於 α 值便接受 H_0。

例 2. 投擲一銅板10次, 且欲檢定

$$H_0 : p = 0.5,$$
$$H_a : p > 0.5,$$

其中 p 爲銅板出現正面之機率。則當觀測到

6個正面, p-值約爲0.377,

7個正面, p-值約爲0.172,

8個正面, p-值約爲0.055,

9個正面, p-值約爲0.0107,

10個正面, p-值約爲0.000977。

當得到8個正面, 若取 $\alpha = 0.10$, 會拒絕 H_0; 若取 $\alpha = 0.05$, 會接受 H_0。

如何決定拒絕域呢? 在同一顯著水準下, 拒絕域往往不唯一。但大家不要忘了還有第二型錯誤: 產品明明不符合規格 (H_0不眞), 檢定結果卻是接受它符合規格 (接受H_0)。這當然不好, 我們也不希望此機率太大。這就產生了*最佳檢定* 的問題。所謂最佳, 就是指在顯著水準不超過一給定的 α 值之下, 第二型錯誤的機率要最小。

假設檢定裡, 仍有相當多的題材。本文只是初步介紹, 給讀者一些基本概念, 細節可參考黃文璋 (2003), 或一般數理統計的書。黃文璋 (2004) 一文也可參考。

最後, 有沒有可能兩型錯誤的機率皆爲0呢? 除了一些 "無聊" 的情況外, 是不會有的。例如投擲銅板10次, 令p表銅板出現正面之機率。對無聊的檢定

$$H_0 : p \leq 1,$$
$$H_a : p > 1,$$

只要取拒絕域爲{正面數> 10}, 可得一兩型錯誤機率皆爲0之檢定。但對檢定

$$H_0 : p = 0.5,$$
$$H_a : p > 0.5,$$

要使第一型錯誤機率爲0不難, 只要取如上之拒絕域即可。但因永不接受H_a, 故當H_a爲眞時, 仍必接受H_0, 故第二型錯誤機率爲1。

5. 結語

假設檢定裡無罪推定的原則務必要掌握。要將現況, 或所欲推翻者置於虛無假設。這不是保守或故步自封, 而是客觀、謹愼, 如此所得到的推論才有說服力, 新的推論也才不會很快又被推翻。假設檢定裡不主張朝令夕改。除非有顯著的差異, 否則寧可維持現況。一件新的產品要能被接受, 其品質 (壽命、效果或價格) 總要顯著地優於舊有的, 否則人還是會習於用舊的。另外, H_0與H_a之挑選也要留意。如果消費者懷疑某飲料容量不足, 想做一檢定, 則要將容量足置於H_0,。反之生產者做品管, 要將容量不足置於H_0。否則可能得到不合理的推論, 見下例。

例 3. 某罐裝飲料標示容量爲330(單位爲 cc), 正負誤差3。某消費者懷疑容量其實只有329。他隨機取36罐做檢驗, 得平均容量330.1。令μ表實際平均容量, 他想檢定

$$H_0 : \mu = 329,$$
$$H_a : \mu > 329。$$

假設容量有常態分佈, 標準差爲3, 則在 H_0 之下, 容量有 $\mathcal{N}(329, 3^2)$ 分佈。36 瓶之平均容量則有 $\mathcal{N}(329, 3^2/36) = \mathcal{N}(329, (1/2)^2)$ 分佈。取 $\alpha = 0.01$, 則拒絕域爲觀測到之平均容量大於 $329 + 2.326 \cdot (1/2) = 330.163$。由於 $330.1 < 330.163$, 所以得到推論爲接受 $H_0 : \mu = 329$。取樣的平均容量大於330, 還要被判定容量不足330, 這樣的推論當然是荒謬的。

正確的作法是, 取

$$H_0 : \mu = 330,$$
$$H_a : \mu < 330。$$

仍取 $\alpha = 0.01$, 則拒絕域爲觀測到之平均容量小於 $330 - 2.326 \cdot (1/2) = 328.837$。由於$330.1 > 328.837$, 故接受 $H_0 : \mu = 330$。

另外, 要提醒讀者一點, 檢定結果若兩個量之間的關係具顯著性, 不表二者有因果關係。聽過蜘蛛聽力的故事吧! 有個學生發現蜘蛛沒有耳朵而有8隻脚, 他懷疑蜘蛛用脚聽聲音。於是他做個實驗, 將蜘蛛放在桌上, 大叫一聲 "爬" , 蜘蛛向前爬動。接著他將蜘蛛的脚全割下, 又大叫一聲 "爬", 蜘蛛沒有爬動。於是他得到蜘蛛用脚聽聲音的推論。這種例子很多, 舉一些如下:

1. 可樂銷售量較大時, 到醫院腸胃科就診之病患增加。多喝可樂會引起腸胃毛病嗎? 可能未必。因夏天天氣熱, 容易吃壞肚子, 此時可樂亦大賣, 但喝可樂不見得會引起腸胃毛病。
2. 體操選手多半較矮, 練體操會使個子變矮嗎? 可能未必, 因當初說不定就是挑選身材較嬌小的人來練體操。

假設檢定裡的接受或拒絕, 只是依據實驗得到數據所做的推論。重做一次會不會得到相反的推論? 當然可能。在這隨機世界, 不論再好的方法, 所做的決策, 總難免有錯。我們只能在所給的條件下, 儘可能減小犯錯機率。保護虛無假設, 寧可錯放而不錯殺。若虛無假設實際是錯的卻接受 (例如產品不符合規格卻讓它通過檢定), 終究有被糾出的一天。那些實際有罪, 而被法庭宣佈無罪釋放的被起訴者, 如果高興地歡呼 "司法還他清白"後, 自此改邪歸正, 那倒無妨, 如果繼續做壞事, 根據機率理論, 無罪的虛無假設, 總有被拒絕的一次。

參考文獻

1. 洪蘭譯 (2005). 恐懼之邦 (Michael Crichton 原著 State of Fear)。遠流出版事業股份有限公司, 台北。
2. 黃文璋 (2003). 數理統計。華泰文化事業股份有限公司, 台北。
3. 黃文璋 (2004). 統計學裡無罪推定的精神。科學發展, 383期 (2004年11月號): 68-73。

—本文作者任教於國立高雄大學應用數學系—