數學傳播 42卷3期, pp. 43-47

由「太陽從西邊升起」談 p 值的意義

連怡斌

最近國際學術界又開始對所謂的 p 值做一連串的批判; 著名的社會學學術期刊《政治分析》宣布在 2018 年起禁用 p 值, 原因之一是 p 值無法對所給定的 model 給予直觀的支持證據。(參考 http://www.sohu.com/a/218689757_119719)。以下我想討論的是, p 值其實是源自最基礎的邏輯論證, 有它相當「直觀」的解釋, 可惜大多統計的教科書只強調其計算和機率分配的意義, 其他領域學者常誤解其意義, 而終究導致這統計學的「命根子」被排斥。

何謂 p 值?

在科學研究上, 我們常面對這樣的判別需求：有某個 (新) 理論是研究者認爲可能是真的 (例如：某種新研發出的降血壓的藥和傳統的藥相比, 有不同的藥效)。姑且稱之爲假說 H_1。而其反面, 也就是和 H_1 互爲補集 (complement) 的說法就稱之爲 H_0;

H_0：新藥和傳統藥的藥效相同,

H_1：新藥和傳統藥的藥效不同。

研究者要如何說服同儕或大衆他認爲的 H_1 是正確的呢? 首先他透過試驗或調查後, 得到某研究結果 A。接著算出, 在 H_0 爲真的前提下, 得到結果 A 的機率, 這就是所謂的 p 值。若 p 值很小, 研究者就推論：「H_0 這前提不太可能爲真」, 所以「H_1 應該是真」。那 p 值要多小才算很小呢? 其實沒有定論, 但大多數接受「小於 0.05 就算很小」這樣的規則。

其實一個直覺的問題馬上跑出來：爲什麼要繞一圈, 靠說「H_0 不太可能爲真」來說明「H_1 應該是真」? 幹嘛不直接以機率衡量 H_1 真實的程度? 由上例可看出一現實的困難：在 H_1 前提下要算出結果 A 的機率相對於 p 值要困難很多, 因爲所謂「藥效不同」有太多種不同法; 藥效好很多跟好一點點都稱爲不同, 在這些不同的情況要算得到結果 A 的機率需其他條件; 但「H_0：相同」下就很直接。不過用計算上的理由來爲 p 值辯護, 可能較難說服其他領域的學者。本文要提的, 是 p 值事實上是源自希臘亞里士多德所發展出來的邏輯辯證方法。

太陽從西邊升起

一些連續劇常會出現類似以下的一幕：丈夫跟妻子發誓：「如果我有做這件事, 太陽就從

西邊升起!」當然, 丈夫有沒有做這件事實在跟太陽沒半點關係; 但他其實真正要表達的是:「因爲太陽不可能從西邊升起, 所以我沒有做這件事!」換成邏輯的論證方式, 第一句話是由論述 p:「我有做這件事」推論到論述 q:「太陽就從西邊升起」; 表示成 $p \to q$ (若 p 則 q)。而學過基礎邏輯的都知道, 和它完全相等的說法是 $\sim q \to \sim p$ (若非 q 則非 p), 也就是第二句話:「因爲太陽不可能從西邊升起, 所以我沒有做這件事!」

現在用統計假說檢定的方式來看, 因爲妻子懷疑「丈夫有做這件事」, 所以我們將妻子的假說訂爲 H_0, 而其相反訂爲 H_1:

H_0: 丈夫有做這件事,

H_1: 丈夫沒有做這件事。

依丈夫的論述, 若 H_0 這個「因」的假設是正確的, 那就會發生「太陽從西邊升起」這樣的果; 但因發生這樣結果不可能 (或者說, 機率爲零), 所以 H_0 的假設不可能對, 換言之那就是 H_1 一定對。這, 不就是我們熟知的「反證法」, 或稱「歸謬證法」或「矛盾證法」嗎?!

說到這裡, 如果您質疑我「說半天還是沒說丈夫到底做了什麼事」, 那就真的搞錯重點了。我要說的是, 上例在假說檢定中, p 值就是:「太陽就從西邊升起」的機率 (太太的懷疑是正確的前提下); 因爲 p 值爲零, 所以 H_0 的假說被推翻而 H_1 成立。所以 p 值根本就是我們日常生活中以別種方式常常在用的推論工具!

零與一

不久前我一位律師朋友在 FB 上有感而發:

「在科學上, 你只能證明某件事存在, 你永遠沒辦法證明一件事不存在; 不存在的意思, 只是現在還沒被發現 $\cdots$」。

我回應他:

「在科學上, 你絕對有辦法證明一件事不存在的, 而且這類論證在日常生活中還常常在用 $\cdots$」(事實上律師用最多!)

舉個高中學過的問題:「最大的正整數存不存在?」

大家都知道它不存在, 但你如何論述? 依照那位律師朋友的說法,「你認爲不存在, 只是你還沒找到而已啊!」幸好我們有上述的反證法:

假設最大的正整數存在, 令它爲 N。由此我們可推論,

(1) $N+1$ 仍爲正整數,

(2) 任何整數加 1 後不可能變小, 所以 $N+1 \geq N$。

(3) 因 N 爲最大的正整數, 所以 $N \geq N+1$。

(4) 由上得到 $N = N + 1$, 移項後 $0 = 1$。

(5) 因爲 0 不可能等於 1, 所以假設不正確, 所以, 其相反假說「最大的正整數不存在」才是正確。

再度用「假說檢定」的觀念來陳述

H_0 : 最大的正整數存在,

H_1 : 最大的正整數不存在。

若 H_0 這個「因」的假設是正確的, 那就會發生「$0 = 1$」這樣的果; 但因發生這樣結果的機率 (即 p 值) 爲零, 所以「存在」的假設不可能對, 換言之那就是 H_1 一定對。

統計歸謬證法

在社會或生物科學領域中, 通常在H_0假設下所得的「果」機率很難爲零, 了不起就是「很小很小」而已。第一例中, 假如丈夫也不是那麼肯定他一定沒做, 那他可能換種說法 :
「如果我有做這件事, 出去就給車子撞死!」
Well, 出門會不會發生車禍這是難講一定不會發生, 但機率肯定很小, 就說是小於萬分之一好了。依統計「假說檢定」的方式論述 :
若 H_0「丈夫有做這件事」的假設是正確的, 就會發生「出去就給車子撞死」這樣的果; 但因發生這樣結果不「太」可能 (機率小於萬分之一), 所以 H_0 的假設不「太」可能對, 換言之那就是 H_1「非常可能」對。相對於前面兩例的「數學反證法」, 這種論述我們可稱之爲「統計反證法」或「統計歸謬證法」; 而 p 值就是「出去就給車子撞死」的機率 (在 H_0 前提下)。

再舉一個相當直觀的例子。莊家提供了一其號稱公正 (正反機率各半) 的銅板和賭客對賭丟銅板的遊戲 : 正面賭客贏, 反面賭客輸。結果連玩 10 次竟然都是反面, 賭客便控訴莊家的銅板不公正。您認爲賭客的控訴合理性多高? 莊家當然可辯稱 : 說不定再多丟十次, 結果會都是正面, 那不就公正了嗎? 將兩人的爭執以假說檢定的方式呈現 :

H_0 : 銅板公正,

H_1 : 銅板不公正。

那有沒有可能銅板是公正 (H_0 正確), 只是賭客運氣不好? 當然有可能。只是, 這麼「背」的運氣發生機率是多少? 簡單計算下爲 $(1/2)^{10} = 1/1024$ (小於千分之一)。也就是說若 H_0 正確, 結果發生了機率很小 (小於千分之一) 的情形, 所以賭客推論 : 銅板很可能不公正。若這樣的推論, 您認爲說服力不夠強 (千分之一也不小啊), 那考慮若賭客連擲 30 次皆反面的情形, 也就是「若銅板是公正, 發生這情形的機率近乎不可能」($(1/2)^{30}$ 小於億分之一), 這樣「證明」銅板有問題的說服力可能大多數人認爲夠強了! 反過來說 : 若賭客連擲3次皆反面, 就指控不公正, 這

樣的說服力就較前者差, 因爲發生這情形的機率「尚有八分之一」($(1/2)^3$)。在這例子中, 不管擲出幾個連續反面, 你永遠無法像證明無最大自然數的例子一樣, 得到絕對爲零的 p 值。

簡而言之, 數學反證法, 是將預期要被推翻的說法放在 H_0, 而其相法的說法放在 H_1; 當 H_0 果真被推翻時, 等同於支持了 H_1 的正確性。而何時 H_0 可被被推翻? 就是

當H_0衍伸或推論出來的結果爲絕不可能發生的事 (如「太陽從西邊升起」, 或「$0 = 1$」) 時, 就果斷推論 H_0 絕對錯誤 (即, H_1 絕對正確)。

而統計反證法和數學反證法唯一的不同, 是考量現實世界的複雜性, 將話說的委婉些:

當 H_0 衍伸或實驗出來的結果爲不太可能發生的事 (如「公正銅板連擲 30 次皆反面」) 時, 就推論 H_0 非常可能是錯誤 (即, H_1 很可能正確)。

而 p 值, 只是 H_0 衍伸或實驗出來之結果其發生的機率! 換句話說, 它是如「太陽從西邊升起」, 或「公正銅板連擲 30 次皆反面」的機率, 而非「丈夫有沒做這件事」或「銅板是否公正」的機率。

而我前面曾提到, 這種反證法, 倒是律師/法官們最常用。對某個命案的被告嫌疑犯, 其審判結果只能是以下兩者其中之一:

H_0: 被告無罪,

H_1: 被告有罪。

檢察官必需提出有力的證據支持 H_1, 才能說服法官判其有罪。原則上, 被告無須證明自己無罪, 這是無罪推定論的原則。假設警方發現命案現場兇手留下的血跡, 經 DNA 比對, 和被告「相當相似」。如何個相似法呢? 這裡講「基因相似度」是沒太大意義的, 因爲隨便找一同性別的白人和黑人, 其基因相似度就高達 99.99%, 那「更相似」是指 99.999% 還是 99.9999%? 較科學的說法, 是反過來講:每一百萬人的基因中, 才會有一個人其基因湊巧和兇手如此相似; 所以如果被告是無辜的 (H_0 正確), 那他/她純粹是運氣差, DNA 和兇手湊巧如此相似的機率爲百萬分之一。也就是 p 值爲百萬分之一。在這種情況下, 法官要有一個閥值:若這「誤判的機率」(即 p 值) 真的夠小, 那就判他有罪吧! 你可以說爲避免冤枉好人, 這閥值要很小, 如千萬分之一或十億分之一才判有罪, 但不能說一定要爲零才能判有罪。因爲除非你就不要相信 DNA 提供的證據, 否則這證據誤判的機率跟擲銅板的例子一樣, 只可能很小很小, 但永遠不會爲零。

不要 p 值之後呢?

醫學期刊或社會學期刊認爲 p 值無法對給定的模式 (也就是假說本身如 H_0 或 H_1) 直接支持的證據, 這也是實情: p 值的確無法直接描述 H_0 或 H_1 爲正確的機率。但問題是, 也沒

有其他有根據的值可以來描述這件事啊! 在很多情況下, H_0 或 H_1 的正確與否, 根本無機率可言, 例如「丈夫有沒做這件事」或「被告是否犯案」是已發生的事, 他要嗎就有, 要嗎就沒有, 機率不是 0 就是 1。如果一定要賦予這假說一個機率, 那可能只有利用貝氏機率的方法。但用貝氏方法的前提, 是要能有足供信賴的 prior information, 如先驗機率等, 但這不在本文範疇, 不在此詳談, 我只能說在很多研究上這是個不易擁有的奢侈品。

至於所謂的其他替代方式, 如「信賴區間」, 如果是熟知統計估計與統計檢定的機率基礎的人, 就瞭解其實兩者是基於同樣的機率架構, 只是換個說法而已。事實上個人感覺,「信賴區間」表面上看起來較直觀, 但實際意義還較 p 值更難解釋些。數年前台灣高中教材放入信賴區間, 就把高中老師們弄得雞飛狗跳, 不知如何教才對! 最後不得已還是將它拿掉。反之, 如果你能接受反證法是個推論邏輯命題正確與否的好方法, 那就比較容易接受「p 值還算是個相對不錯的評估工具」。但重點是, 統計教師們要能將 p 值跟這常用的邏輯脈絡做連結, 說明當初紐曼 (Neyman) 及皮爾生 (Pearson) 在 1933 提出 p 值時的觀點, 是相當符合人類思考方式的。如此或有可能說服其他領域的學界重新接受 p 值的價值, 並減少對它的誤解。

—本文作者爲國立彰化師範大學統計資訊所及數學系教授—