淺談 p 值

王博賢 副統計分析師
陳錦華　副教授

p 值(p-value)一直以來常用於**"判斷統計上是否顯著"**的一個重要指標。在檢定上，大家都知道 p 值小於顯著水準(門檻值常設定為 0.05)，我們就拒絕虛無假設，若是應用於檢定迴歸係數是否顯著時，則表示這個效應(effect)顯著。簡單、易懂、好記，然而這就是真理嗎?

美國協會統計協會(American Statistical Association, ASA) 在 2016 年的聲明中提到一段對話[1]:

Q: Why do so many colleges and grad schools teach p = 0.05?

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people still use p = 0.05?

A: Because that's what they were taught in college or grad school.

為什麼學校都在教 p 值 = 0.05？ 因為科學雜誌都在用。為什麼大家都在用 p 值 = 0.05？ 因為學校都在教。然而因為這個樣子 p 值 已慢慢地被濫用了，不少人只要看到 p 值 < 0.05 就做出結論。不過也有越來越多學者意識到它的缺點和濫用的可能性，甚至 Basic and Applied Social Psychology 這個心理學期刊，在 2015 年開始，已決定往後投稿的文章，不能只使用 p 值 ，而需提供信賴區間的訊息，以做為讀者對於數據更深的理解。

最簡單能想到解決這個問題的方法就是調降 p 值 的門檻值。像是 Nature Human Behaviour 期刊中 2017 年 7 月份預印本，有 72 學者針對 p 值 提出不同的看法及意見[2]：

> 他們期望可以推出一個重大制度的改變，在社會科學和生物醫學的文章中，p 值的門檻值不再是 0.05，而是 0.005。在所有的科學都設定這樣的門檻是過於極端，但這是個時機，要求研究者證明他們做了什麼？科學就能有進步。然而降低門檻值可能導致 'file-drawer problem' 更嚴重，這也就是我們在整合分析(meta-analysis)中所說的'publication bias(發表偏差)'，有顯著的結果傾向被發表於期刊上，我們也只接受到部份的訊息，另一部份的研究結果，若不顯著則不易刊登於期刊上，間接導致我們做出更多不正確的決定。大家皆表示，在研究前(實驗開始前)，對於 p 值而言研究者應選擇適當的門檻值，此值應根據研究之因子個數、此門檻值應被評估並於研究前登錄，並有一審查機制對於科學方法及分析進行評審，以確保過程之正確性。令人驚訝的是，他們進了調查(2017/07)，" 你認為：p 值的門檻應再低一些？" 有 69%的人，認同意。

美國統計協會執行主任 Ronald Wasserstein 就說“我們不應該驚訝，這世界沒有一個魔術數字”。他認為為科學設下一個特定標準的門檻，是在傷害科學。

其實因為最近 p 值 這個問題，吵得滿火熱的，因此 ASA 為此發表了一篇對 p 值 的聲明，讓我們來看看他們的聲明內容。

1. **P-values can indicate how incompatible the data are with a specified statistical model.**

p 值 可以提供我們判斷數據與特定模型之間的不兼容性，也是代表樣本數據和虛無假設下之差異的證據大小。在用於計算 p 值 的基本假設不變之下，這種不兼容性可以用來解釋或提供我們對背後的基本假設及虛無假設的懷疑。要注意的是這邊提的**特定模型是虛無假設為真情況下得模型**。

2. **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**

p值 不能拿來計算研究假設為真的機率，或者是算這筆資料單獨被生成的機率。而是用來對資料證據與虛無假設之間的關係做說明，是要證明虛無假設是錯的。而顯著水準(門檻值)是：若虛無假設為真，有 5%的機率，資料會和虛無假設不合。

3. **Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**

科學結論和商業或政治決策不應該單單靠 p 值 是否通過一個指定的門檻，雖然做出“yse or no”這個決定是重要的，但這也不表示單靠 p 值 就可以做決定，而是應該要引用背後的理論，研究的設計、測量方法及一些外部證據來協助做出結論。不然就會像 p 值 < 0.05，一樣導致科學研究變得扭曲。

4. **Proper inference requires full reporting and transparency.**

做出適當的推論需要充分的報告以及透明度，我們都知道 p 值 其實是容易可以操控的，其中最常見的是“Cherry-picking”的問題[3]。例如：每次多一個變數或少一個，增加一點資料或少一點資料，這樣慢慢去測試直到顯著為止，檢定的次數早就不止一次，雖然每次檢定都把顯著水準設為 0.05，其實這麼早已把我們犯錯的機會變高了，也就是說型 I 錯誤早就高於 0.05。因此公開透明是很重要的，就有人建議應該要採取事先登記(pre-registration)，研究前研究者應該把整個研究計畫做個事先登記，且之後不能在改動，如果之後發表的文章跟事先登記的不一樣，那就可以懷疑他的正確性。

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

p 值 或統計上的顯著，並不能直接代表效果的大小、或結果的重要性。較小的 p 值代表有較強的樣本證據顯示，資料來源和虛無假設下的條件很不相同，因而較小的 p 值不一定意味著有較大效應(effect)影響，只要樣本數夠大、測量精度(measurement precision)夠高(即標準誤較小)，也能產生較小的 p 值。反之，樣本數小或測量精度不夠高，導致 p 值 不正確而產生較大的值也是有可能的。另外，統計上的顯著差異，不代表實際運用上之顯著(醫學上、臨床上、商用上)，在應用及解釋上要特別注意。

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

p 值 本身在推論一個模型或假設是否正確上，其實並不太好。我們不應該把計算 p 值 當成分析的最後一步，應該配合其他的方法，例如信賴區間估計等，才能得到較為精準的統計推論。

ASA 提供補充或取代 p 值 的方法:

- Confidence, credibility, or prediction intervals
- Bayesian methods.
- Likelihood ratios or Bayes Factors.
- Decision-theoretic modeling and false discovery rates.

以上方法雖然需要進一步的假設，但他們可能可以更直接地顯示出效果的大小或假設的模型是否正確。

最後 ASA 強調良好的統計分析應該建立在好的實驗設計及好的品行上，藉由各種數據，及圖形進行適當且具有邏輯的推論。世上並沒有一個單一指標可以取代科學推理。

參考資料:

1. Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose.
2. 'One-size-fits-all' threshold for *P* values under fire**.** http://www.nature.com/news/one-size-fits-all-threshold-for-p-values-under-fire-1.22625#/ref-link-2
3. 林澤民(2016 / 10 / 01). 看電影學統計：p 值的陷阱. 社會科學論叢；10 卷 2 期 。