社會科學前沿課題論壇

看電影學統計：p 值的陷阱*

The Pitfalls of p-Values

奧斯汀德州大學政府系副教授　林澤民
Tse-Min Lin
Associate Professor, Department of Government,
University of Texas at Austin

* 本篇為 105 年 6 月 6 日12:00-14:00 論壇講座內容。

看電影學統計：p 值的陷阱

林澤民
奧斯汀德州大學政府系副教授

院長、陳老師，各位老師、各位同學，今天很榮幸能夠到政大來，和大家分享一個十分重要的課題。我今年回來，今天是第六個演講，六月中之前還有兩個，一共八個，其中四個是談賽局理論，四個是談 p 值的問題。賽局理論的部分，題目都不一樣，譬如我在政大公行系講賽局理論在公行方面的應用，而我第一個演講在台大地理系，談賽局理論在電影裡的應用。我在台大總共講了三部電影，第一部是「史密斯任務」，講男女關係、夫妻關係；第二部是「少年 pi 的奇幻漂流」，講少年和老虎對峙的重複性賽局；第三部電影是最新的電影：「刺客聶隱娘」，講國際關係賽局。今天談的當然是不一樣的題目，雖然它是一個很重要、很嚴肅的題目，但我希望大家可以輕鬆一點，所以也要放兩部電影片段給大家看，一部是「玉蘭花」，另一部則是「班傑明的奇幻旅程」，這兩部電影都有助於我們來瞭解今天要談論的主題：p 值的陷阱。

科學的統計學危機

爲什麼要談論 p 值的問題？因爲在近十多年來，不只是政治學界，而是很多學門，特別是在科學領域，有很多文章討論傳統統計檢定方法、尤其是 p 值統計檢定的問題，甚至有位很有名的統計學者，Andrew Gelman 寫了篇文章，叫作 The Statistical Crisis in Science──「科學的統計學危機」，說是危機一點都不言過其實。這就是爲何我說：今天要討論的其實是很嚴肅的問題。

投影片上這些論點，大部分是說我們在傳統統計檢定的執行上，對 p 值有各種誤解跟誤用。現在很多人談到「p 值的危險」、「p 值的陷阱」、「p

值的誤用」、還有「p 值的誤解」。甚至有些學術期刊，也開始改變他們的編輯政策。像這本叫作 Basic and Applied Social Psychology 的心理學期刊，已經決定以後文章都不能使用 p 值，大家能夠想像嗎？我們作計量研究，都是用 p 值，各位一直用，在學界用了將近一百年，現在卻說不能用。甚至有些文章，說從前根據 p 值檢定做出來的研究成果都是錯的，有人更宣告 p 值已經死了。所以這是一個很嚴重的問題。在這本期刊做出此決定後，美國統計學會（ASA）有一個回應，表示對於 p 值的問題，其實也沒這麼嚴重，大部分是誤解跟誤用所造成，只要避免誤解與誤用就好。可是在今年，ASA 眞的就發表了正式聲明，聲明裡面提出幾點，也是我今天要討論的主要內容，包括 p 值的眞正的意義，以及大家如何誤用，換句話說就是：p 值到底是什麼？它又不是什麼？（圖 1）今天除了會深入探討這些議題之外，也請特別注意聲明的第三點提到：科學的結論，還有在商業上、政策上的決策，不應只靠 p 值來決定。大家就應該瞭解這問題影響有多大、多嚴重！

ASA關於P值的聲明(2016)

- P-values can indicate how incompatible the data are with a specified statistical model.
- 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.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

■圖 1　ASA 關於 p 值的聲明

我舉個例子，最近在台灣，大家都知道我們中研院翁院長涉入了浩鼎案，浩鼎案之所以出問題，就是因爲解盲以後，發現實驗的結果不顯著。我今天不想評論浩鼎案，但就我的瞭解，食藥署、或者美國的 FDA，他們在批准一項新藥時，一定要看實驗的結果，而且實驗結果必須在統計上要顯著。可是

ASA 卻告訴我們說，決策不該只根據統計的顯著性，大家就可想像這影響會有多大。甚至有其他這裡沒有列出來的文章，提到爲何我們使用的各種藥物，都是經過這麼嚴格的 p 值檢定出來、具有顯著性，可是在眞正臨床上，卻不見得很有用。其實很多對 p 值的質疑，都是從這裡出來的。

有關 p 值的討論，其實並非由政治學門，而是從生命科學、例如醫學等領域所產生的。ASA 聲明的第四點說：正確的統計推論，必須要「full reporting and transparency」，這是什麼意思呢？這是說：不但要報告 p 值顯著的研究結果，也要報告 p 值不顯著的研究結果。但傳統方法最大的問題是：研究結果不顯著，通通都沒有報告。在英文有個詞叫 cherry-picking，摘櫻桃。什麼叫摘櫻桃？摘水果，水果熟的才摘，把熟的水果送到水果攤上，大家在水果攤上看到的水果，都是漂亮的水果，其實有很多糟糕的水果都不見了。我們在統計上也是，大家看到的都是顯著的結果，不顯著的結果沒有人看到。可是在過程中，研究者因爲結果必須顯著，期刊才會刊登、新藥才會被批准，所以盡量想要擠出顯著的結果，這之中會出現一個很重大的問題：如果我們作了 20 個研究，這 20 個研究裡面，虛無假設都是對的，單獨的研究結果應該是不顯著。可是當我們作了 20 個統計檢定時，最少有一個結果顯著的或然率其實很高。雖然犯第一類型錯誤的或然率都控制在 0.05，可是 20 個裡面最少有一個顯著的，或然率就不是 0.05，大概是 0.64。如果就報告這個顯著結果，這就是 cherry-picking。ASA 給的建議是：實驗者必須要 full reporting and transparency，就是一個研究假如作了 20 個模型的檢定，最好 20 個模型通通報告，不能只報告顯著的模型。ASA 這個聲明是今天要討論的主要內容。

p 值是什麼？

p 值是什麼？我想在座有很多專家比我都懂，但是也有一些同學在場，所以還是稍微解釋一下。p 值是由 Ronald Fisher 在 1920 年代發展出來的，已將近一百年。p 值檢定最開始，是檢定在一個 model 之下，實驗出來的 data 跟 model 到底吻合不吻合。這個被檢定的 model，我們把它叫做虛無假設（null hypothesis），一般情況下，這個被檢定的 model，是假設實驗並無系統性效

應的，即效應是零，或是隨機狀態。在這個虛無假設之下，得到一個統計值，然後要算獲得這麼大（或這麼小）的統計值的機率有多少，這個或機率就是 p 值。

舉一個例子，比如說研究 ESP——超感官知覺——時會用到比例（proportion）這個統計值。我們用大寫的 P 來代表比例，不要跟小寫的「p 值」的 p 混淆。在 p 值的爭論裡，有一篇研究 ESP 的心理學文章被批評得很厲害。文章中提到了一個實驗，讓各種圖片隨機出現在螢幕的左邊或者右邊，然後讓受測者來猜圖片會出現在哪邊。我們知道如果受測者的猜測也是隨機的，也就是沒有 ESP 的效應，則猜對的或然率應該是一半一半，算比例應該是差不多 $P = 0.5$，這裡比例 $P = 0.5$ 就是我們的虛無假設。但這個實驗——實驗者是一位知名心理學教授——他讓受測者用各種意志集中、力量集中的辦法，仔細地猜會出現在左邊還是右邊。結果發現，對於某種類型的圖片——不是所有圖片，而是對於某些類型的圖片，特別是色情圖片——受測者猜對的比例，高達 53.1%，而且在統計上是顯著的。所以結論就是：有 ESP，有超感官知覺。

這裡 p 值可以這樣算：就是先做一個比例 P 的 sampling distribution——抽樣分配。如果虛無假設是對的，平均來講，$P = 0.5$。0.5 就是 P 的抽樣分配中間這一點，這個比例就是我們的虛無假設。在受測者隨機猜測的情況之下，P 應該大約是 0.5 的。可是假如真正得到的 P 是 0.531，抽樣分配告訴我們：如果虛無假設是對的，亦即如果沒有任何超自然的力量，沒有 ESP 存在，大家只是這樣隨機猜測的話，則猜對的比例大於或者等於 0.531 的機率，可以由抽樣分配右尾的這個面積來算。作單尾檢定，這面積就是所謂的 p 值。如果作雙尾檢定的話，這值還要乘以 2。以上就是我們傳統講的 p 值的概念。

我們得到 p 值以後，要作統計檢定。我們相約成俗地設定一個顯著水準，叫做α，α 通常都是 0.05，有時候大家會嚴格一點用 0.01，比較不嚴格則用 0.10。如果我們的 $\alpha = 0.05$，則若 $p < 0.05$，我們就可以拒絕虛無假設，並宣稱這個檢定在統計上是顯著的，否則檢定就不顯著，這是傳統的 p 值檢定方法。如果統計上顯著的話，我們就認爲得到實驗結果的機會很小，所以就不接受虛無假設。爲什麼說 p 值很小，就不接受虛無假設？我個人的猜想，這是依據命題邏輯中，以否定後件來否定前件的推論，拉丁文稱作 *modus tollens*，意思

是以否定來否定的方法，也就是從「若 P 則 Q」和「非 Q」導出「非 P」的推論，這相信大家都知道。P 值檢定的邏輯是一種有或然性的 *modus tollens*，是 *probabilistic modus tollens*。「若 H_0 爲眞，則 p 值檢定顯著的機率很小，只有 0.05」，現在 p 值檢定顯著了，所以我們否定 H_0。但是命題邏輯的 *modus tollens*，「若 P 則 Q」是沒有或然性、沒有任何誤差的餘地的。「若 H_0 爲眞，則 p 值檢定不可能顯著」，這樣 p 值檢定顯著時，你可以否定 H_0，大家對此都不會有爭議。問題是假如容許或然性，這樣的推論方法還是對的嗎？舉一個例子：「若大樂透的開獎機制是完全隨機的，則每注中頭獎的機率很小，只有 1/13,980,000」，現在你中獎了，你能推論說大樂透開獎的機制不是隨機的嗎？p 值的問題，便是在於我們能不能夠因爲 p 值很小，小到可能性很低，我們就用否定後件的方法來否定前件。我們用命題邏輯來作統計推論，但其實我們的推論方法跟命題邏輯卻不完全一樣，因爲我們的 α 絕對不可能是零，如果 α 是零的話，就不是統計了。

再來就是看電影時間，電影很有趣，可以幫助我們瞭解什麼是 p 值，也可以再接著討論爲什麼用 p 值來作統計推論會有錯。這部電影叫做「玉蘭花」，是1999年的電影，已經很舊了，可能在座年輕的朋友就沒看過。網路上在Youtube有這一段，請大家觀賞。（電影「玉蘭花」短片連結：https://www.youtube.com/watch?v=Ec51smvcsDY）

相信大家應該都看得懂這短片的用意。玉蘭花這部電影，雖然裡面有講一些髒話，但是其實是一部傳教的影片。它的推論方式，其實就是我剛剛講的 p 值的推論方式，它有一個虛無假設，就是說事情發生沒有什麼超自然的力量在作用，都是隨機發生的，是 by chance，不是 by design，可是它發生了，竟然有這麼巧合的事情。大家可以想一下，如果事情的發生都是 by chance，都是隨機的，那麼像這種事件發生的機率有多少？很小很小，0.0...01，幾乎不可能發生。所以假如是隨機發生的，就幾乎不可能發生，可是它發生了，我們就以否定後件來否定前件，推論虛無假設——by chance 的這個假設–是不對的。既然不是 by chance，它是什麼？就是 by design，是設計出來的。這是基督教的一種論證上帝創造世界的方法。在美國，有些學區還在爭論，生物是創造的還是進化的？創造論的主張者都會用這樣的論證，說你看我們人體，它是

這麼複雜的一個系統，這種系統可能是隨機發生的嗎？若是隨機發生，機率有多少？是 0.0...01，所以它不可能是隨機發生，因此是創造的。這個理論叫做 intelligent design——智慧的設計——即我們這個世界都是上帝創造、是上帝很有智慧地依照藍圖設計出來的。我今天也不想爭辯這種推論對不對，我只是舉例來說明這種推論的邏輯。

p 值不是什麼？

我本來放這部電影都是爲了在教學上解釋 p 值的概念，可是後來當我注意到對於 p 值的爭議之後，覺得其實這一部電影也可以用來幫我們瞭解爲什麼用 p 值來做統計推論有可能是錯的。

下面這個表是大家都熟悉的（圖 2）。我們可以用這個表來呈現有關虛無假設是對或者不對，是被拒絕或者被接受的四種可能性，其中兩種是作出錯誤統計推論的情況。第一個情況，虛無假設是對的，但統計檢定是顯著的，因此虛無假設被推翻了。這種情況叫做 Type I error，我們保留了 $\alpha = 0.05$ 的機率容許它存在。第二個情況，如果虛無假設是錯誤的，但統計檢定不顯著，所以它沒有被推翻，這個情況叫做 Type II error。Type II error 剛學統計的同學可能不太瞭解，因爲我們通常都不會很清楚地去計算它的機率——所謂 β。這個 β 跟 α 不一樣，不是你可以用相約成俗的方法來訂定，而是會受到若干因素的影響。簡單來講，在一定的顯著水準 α 之下，β 跟樣本大小有關係；樣本太小的話，β 會比較大。另外它跟實驗效應的大小也有關係，如果效應很小的話，β 也會比較大。換句話說，如果虛無假設跟研究假設的距離比較小的話，β 會比較大。可是一般人不會去計算 β，因爲還沒做實驗之前，其實也不知道實驗的效應有多少。儘管如此，β 是可以計算的。算出來了，則我們拒絕錯誤虛無假設，而作出正確統計推論的機率是 $1-\beta$，這 $1-\beta$ 我們就把它叫做「檢定的強度」——the power of the test——我待會兒會用到這個名詞。依此定義，β 愈小的話，power 就愈大。用醫學的術語來說，α，Type I error 的機率，就是僞陽性的機率，而 β，Type II error 的機率，就是僞陰性的機率。

Type I and Type II Errors

		H_0 Is	
		True	False
H_0 Is	Rejected (Test significant)	Type I Error (Probability=α)	Correct Inference (Power=1-β)
	Not Rejected (Test not significant)	Correct Inference	Type II Error (Probability=β)

- α=Pr(Type I Error)=Pr(H_0 Rejected | H_0 True)=僞陽性機率
- β=Pr(Type II Error)=Pr(H_0 Not Rejected | H_0 False)=僞陰性機率

■圖 2　Type I 與 Type II error

我們可以開始討論：傳統用 p 值來作統計檢定方式，爲什麼有問題？剛剛 ASA 的聲明說：p 值 do not measure the probability that the studied hypothesis is true。p 值告訴你：如果虛無假設是對的，你「觀察到資料」的機率有多少，但它並沒有告訴你「虛無假設是對的」的機率有多少，或「研究假設是對的」的機率有多少。這是很不一樣的：前者是 data 的機率，後者是 model 的機率。進一步說明，p 值是在虛無假設爲眞的條件之下，你觀察到和你所觀察到的統計值一般大小（或更大／更小）的機率。但我們作檢定的時候，我們是看 p 值是不是小於你的統計水準 α，如果 $p < \alpha$，我們就說統計是顯著的。換句話說，如果虛無假設爲眞，那麼你的檢定是顯著的機率是 $\alpha = 0.05$。但這其實不是我們作研究最想回答的問題；這個機率只告訴我們，如果你的虛無假設爲眞，有 5% 的機率，data 會跟它不合，但它沒有告訴我們虛無假設這個 model 爲眞的機率有多少，而這才是我們應該問的問題。所以我們應該反過來問，如果你統計檢定是顯著的，在此條件之下，「虛無假設是對的」的機率有多少？如果我們把關於 data 這個僞陽性的機率記作 $\alpha = \Pr(Test = +|H0)$，大家可以看出這個關於 model 的機率其實是它倒反過來的：$\Pr(H_0|\ Test = +)$，所以我把它稱作「僞陽性的反機率」。這兩個機率原則上不會相等；只有在 $\alpha = 0$ 的時候，兩者才都是零而相等。

譬如今天你去健康檢查，醫生給你做很多篩檢，如果篩檢結果是陽性，其實先不要怕，因爲你應該要問，如果篩檢出來是陽性，那麼你眞正並沒有病的

機率是多少？也就是僞陽性的反機率有多少？大家可能會很驚訝，僞陽性的反機率通常都很高，但是這個機率，p 値並沒有告訴你。所以必須要去算在檢定是陽性的條件下，結果是一種僞陽性的反機率；這就必須要用「貝式定理」來算。

雖然在座有很多可能比我更高明的貝氏統計學家，但我還是要說明一下貝式定理。先舉一個我終身難忘的例子，剛剛陳老師說我是台大電機系畢業的，我在電機系的時候修過機率這一門課。我記得當時的期中考，老師出了一個題目，說我口袋裡面有三個銅板，其中有一個銅板是有偏差的銅板，偏差的銅板它得到正面的機率是 1/3 ── 不是 1/2 ── 而得到反面的機率是 2/3。考題問：現在我隨機從口袋裡面掏出一個銅板，這個銅板是那個偏差銅板的機率是多少？很簡單大家不要想太多，1/3 嘛。可是我現在拿銅板丟了一下，出現的是正面，我再問你這個銅板是那個偏差銅板的機率是多少？我不期望大家立刻回答，因爲要用貝式定理來算，當你獲得新的資訊的時候，新的資訊會更新原來的機率。這裡我也沒有時間詳細告訴大家怎麼算，但是可以告訴大家，結果是 1/4。如果我丟擲銅板，它得到了正面，它是偏差銅板的機率變成只有 1/4。這是因爲偏差銅板出現正面的機率，比正常銅板要小，所以出現正面的話，它相對來講就比較不太可能是偏差的銅板，所以機率會比原來的 1/3 小些，只有 1/4（大家可以想像如果偏差銅板出現正面的機率是 0，而丟擲得到正面，則此銅板是偏差銅板的機率當然是 0）。原來所知的「1/3 的機率是偏差銅板、2/3 的機率是正常銅板」這個機率分配在貝氏定理中叫做先驗機率（prior probability）。大家要建立這個概念，即是還沒觀察到數據之前，對於模型的機率有一些估計，這些估計就叫做先驗機率。至於觀察到數據之後所更新的模型機率，1/4 和 3/4，這個機率分配叫做後驗機率（posterior probability），也就是前面所說的反機率（inverse probability）。

我們再來看另外一個跟統計檢定問題非常接近的例子。可以用剛剛身體檢查的例子，但我這裡用美國職棒大聯盟對球員的藥物檢查爲例，也許比較有趣。這裡假設大約有 6% 的美國 MLB 的球員使用 PED（performance enhancing drugs），這是一種可以增強體能表現的藥物，是類固醇之類的藥物。這個估計數字可能是眞的，是我從網頁上抓下來的。這邊的 6% 即爲我前面說的先

驗機率：隨機選出一個球員，則他有使用 PED 的機率是 0.06，沒有使用 PED 的機率是 0.94。現在大聯盟的球員都要經過藥檢；舉大家熟知的火箭人 Roger Clemens 爲例。他也是我心目中的棒球英雄，他被檢定有陽性的反應。爲了方便起見，假設藥檢的準確度是 95%。所謂準確度 95% 的定義是：如果一個球員有使用藥物，他被檢定出來呈陽性反應的機率是 0.95；如果一個球員沒有使用藥物，他被檢定出來呈陰性反應的機率也是 0.95。也就是我假設兩種誤差類型的機率 α 跟 β 都是 0.05。在這假設之下，使用貝式定理來計算，當球員被篩檢得到的結果是陽性，但他並不是 PED 使用者的後驗機率或反機率，其實高達 0.45。大家可以從圖 3 看到貝氏定理如何可以算出這個機率。

如何用貝式定理算僞陽性之反機率：
以職業運動「體能增強藥物」(PED)檢測爲例

		Player Is a PED	
		Nonuser	User
Player Tests	Positive (+)	**Pr(+|Nonuser)=.05**	**Pr(+|User)=.95**
	Negative (-)	**Pr(-|Nonuser)=.95**	**Pr(-|User)=.05**
Prior Probabilities		**.94**	**.06**

$$
\begin{aligned}
&\Pr(Player = Nonuser \mid Test = +)\\
&= \frac{\Pr(Player = Nonuser, Test = +)}{\Pr(Test = +)}\\
&= \frac{\Pr(Player = Nonuser, Test = +)}{\Pr(Player = Nonuser, Test = +) + \Pr(Player = User, Test = +)}\\
&= \frac{\Pr(Player = Nonuser)\Pr(Test = + \mid Player = Nonuser)}{\Pr(Player = Nonuser)\Pr(Test = + \mid Player = Nonuser) + \Pr(Player = User)\Pr(Test = + \mid Player = User)}\\
&= \frac{(.94)(.05)}{(.94)(.05) + (.06)(.95)} = .45
\end{aligned}
$$

■圖 3　如何用貝氏定理算偽陽性之反機率：以職業運動「體能增強藥物」（PED）檢測為例

使用貝式定理算出來的結果大家應該會覺得很詫異，因爲我們藥物篩檢的工具應該是很準確的，0.95 在我們想像中應該是很準確的，我們認爲說我們錯誤的可能性只有 5%，其實不然。檢定是陽性，但其實僞陽性的反機率可以高達 45%！所以雖然我不是醫學專家，不過大家健康檢查，如果醫生說，你的檢查結果呈現陽性反應，大家先不要慌張，你要先問一下醫生檢驗的準確度

大概有多少，如果一個眞正有這種病的人來檢定，呈現僞陽性的機率有多少？如果一個沒有病的人來檢定，呈現僞陰性的機率有多少，然後再問他先驗機率大概有多少？然後自己用貝氏定理去算一下僞陽性的反機率。醫學上很多疾病，在所有人口裡面，得病的比例通常很小的。也就是說，得病的先驗機率通常都很小，所以僞陽性的反機率會很大。

現在換成了統計檢定，看下圖的表格（圖 4）。這表格跟圖 3 的表格很像，只是把內容改成了圖 2 的內容：虛無假設是眞的、或是假的，然後統計檢定是顯著、或是不顯著的。然後再加上一行先驗機率，就是「虛無假設是對的」的先驗機率有多少、「虛無假設是錯的」的先驗機率有多少，都用符號來代替數目。我們可以用貝式理得到一個公式，顯示僞陽性的反機率是統計水準 α、檢定強度（power = 1－β）、和研究假設之先驗機率（$P(H_A)$）的函數。α 跟檢定強度都沒問題，但公式裡頭用到先驗機率。你會問：在統計檢定裡面，先驗機率是什麼？

如何用貝式定理算僞陽性之反機率：
P值顯著時H_0爲真之機率Pr(H_0=True|Test=+)

		H_0 Is	
		True (H_A not accepted)	False (H_A accepted)
H_0 Is	Rejected (Test +)	**Pr(+|H_0)=α**	**Pr(+|H_A)=Power =1-β**
	Not Rejected (Test -)	**Pr(-|H_0)=1-α**	**Pr(-|H_A)=β**
Prior Probabilities		**Pr(H_0)**	**Pr(H_A)**

$$
\begin{aligned}
&\Pr(H_0 = True \mid Test = +) \\
&= \frac{\Pr(H_0 = True, Test = +)}{\Pr(Test = +)} \\
&= \frac{\Pr(H_0 = True, Test = +)}{\Pr(H_0 = True, Test = +) + \Pr(H_0 = False, Test = +)} \\
&= \frac{\Pr(H_0 = True)\Pr(Test = + \mid H_0 = True)}{\Pr(H_0 = True)\Pr(Test = + \mid H_0 = True) + \Pr(H_0 = False)\Pr(Test = + \mid H_0 = False)} \\
&= \frac{\Pr(H_0)\alpha}{\Pr(H_0)\alpha + \Pr(H_A)(1-\beta)} \\
&= \frac{(1-\Pr(H_A))\alpha}{(1-\Pr(H_A))\alpha + \Pr(H_A)Power}
\end{aligned}
$$

■圖 4　如何用貝氏定理算偽陽性之反機率：p 值顯著時 H_0 為真之機率

在此我必須要稍微說明一下，先驗機率，以淺白的話來講，跟你的理論有

關係，怎麼說呢？如同剛剛提到 ESP 的實驗，好像只要就這樣用力去猜，你猜對的可能性就會比較高。發表這樣子的實驗報告，我們有沒有辦法告訴讀者，當受測者這樣皺著眉頭去想的時候，到底是什麼樣的一個因果機制，能夠去猜到圖片是出現在左邊還是右邊。

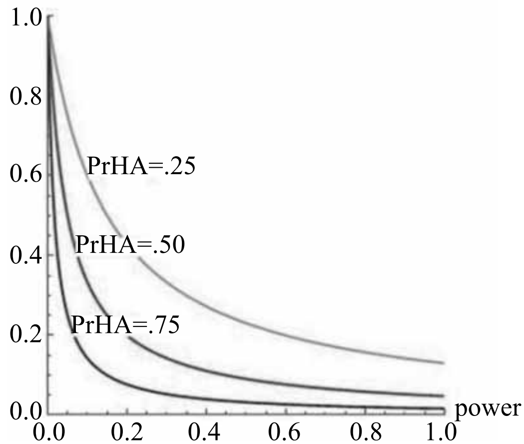
一般來說這種 ESP 的實驗，是沒有這種理論的，是在完全沒有理論的條件之下來做實驗。在此情況之下，我們可以說，此研究假設的先驗機率很小很小。當然我們作政治學的研究就不一樣，我們可能引用很多前人的著作，都有一個文獻回顧，我們也引用很多理論，然後我們說：我們的研究假設是很有可能展的。假如你有很好的理論，你的研究假設的先驗機率就會比較高，在這種情況之下，問題會比較小。但是還有一個問題，就是如果從文獻裡面來建立理論，來判定你的研究假設的先驗機率有多少，問題出在於：通常文獻回顧是從學術期刊裡面得來，而現在所有的學術期刊，發表的都是顯著的結果，不顯著的結果通通都沒有發表，從學術期刊上來判斷研究假設的先驗機率有多少，這樣的判斷是有偏差的。這是我今天要講的第二個問題，現在先繼續討論偽陽性反機率的問題。

現在要詳細討論影響偽陽性反機率的因素，就是影響到「統計檢定是顯著的條件之下，虛無假設爲眞」這一個機率的因素。這裡再重複一下，我們一般瞭解的統計推論，奠基於虛無假設爲眞時，p 值顯著的機率，也就是偽陽性的機率被控制在 α 之內：$\Pr(\text{Test} = +|H_0) = \Pr(p < \alpha|H_0) = \alpha$。但我們現在要反過來問的是：統計檢定是顯著的情況下，H_0 爲眞的機率，也就是偽陽性的反機率：$\Pr(H_0| \text{ Test} = +) = \Pr(H_0| \ p < \alpha)$，這好比篩檢結果爲陽性、但其實球員並未使用 PED、患者其實無病的機率。如果 α 等於零，可以很清楚的發現，這兩個機率是一樣的，都是零；但 α 不等於零的時候，它們就不一樣。由下圖來看，偽陽性的反機率跟先驗機率——研究假設的先驗機率——以及檢驗的強度有關（圖 5、6）。看圖可以得知，power 愈大，還有先驗機率愈大的話，偽陽性的反機率就愈小。可是當 power 愈小的時候，還有先驗機率愈小的時候，偽陽性的反機率就愈大。

偽陽性之反機率Pr(H_0 = True|Test=+)

作爲檢定強度(Power of Test)之函數

・α = 0.5 & Pr(H_A) = .25/.50/.75(Pr(H_0)) = .75/.50/.25

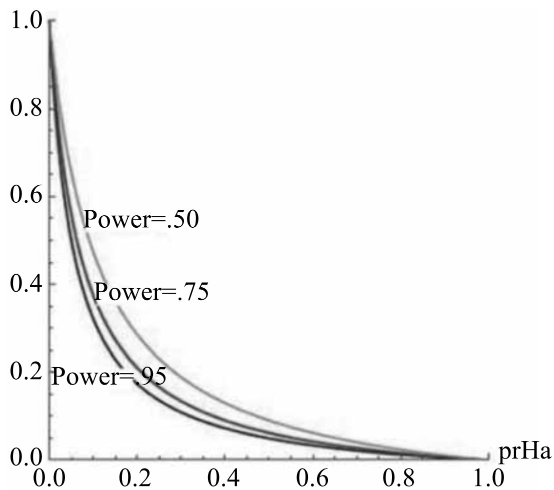


■圖 5　偽陽性之反機率 Pr(H_0 = True | Test = +) 作為檢定強度之函數

偽陽性之反機率Pr(H_0 = True|Test=+)

作爲研究假設先驗機率Pr(H_A)之函數

・α = 0.5 & β = .05/.25/.50(Power = .95/.75/.50)



■圖 6　偽陽性之反機率 Pr(H_0 = True | Test = +) 作為研究假設先驗機率之函數

我做了一個表，列出研究假設的先驗機率，從最小排列到最大，可以看到在不同檢定強度之下，僞陽性的反機率是多少（圖 7）。它可以高到近乎 1.00。換句話說，研究假設的先驗機率如果很小很小，則即使 p 值檢定顯著，但虛無假設仍然爲眞的機率其實還是很大很大的。如果研究假設的先驗機率是 0.5——你事先也許不知道哪一個是對的，你假設是 0.5，就像丟銅板一樣，此時，僞陽性的反機率才是 0.05，才跟 α 一樣。也就是說，研究假設的先驗機率必須要高於 0.5，僞陽性的反機率才會小於 0.05。可是假如你的研究假設，譬如剛剛提到的 ESP 研究，這種實驗沒有什麼理論、沒有什麼因果關係，然後你就去做了一個統計分析。換句話說這個研究假設的先驗機率可能很低，此時僞陽性的反機率其實是很高的。圖 7 第一欄是假設 power 爲 0.95，如果 power 低一點到 0.75 呢？如果是 0.50 呢？我們可以看到其實結果差不多。當然 power 愈低，問題會愈嚴重，但其實差不多，當你的先驗機率是 0.5 的時候，原來是 0.05，現在是 0.09，所以差別不是特別大。原則上，power 對於僞陽性反機率的作用不是那麼強，作用強的是 prior，即是研究假設的先驗機率。

研究假設之先驗機率**Pr(H_A)**的重要性

- α=.05 & β=.05/.25/.50 (**Power=.95/.75/.50**)

Pr(H_A)	Pr(H_0|+)	Pr(H_0|+)	Pr(H_0|+)
0.001	0.98	0.99	0.99
0.01	0.84	0.87	0.91
0.1	0.32	0.38	0.47
0.2	0.17	0.21	0.29
0.3	0.11	0.13	0.19
0.4	0.07	0.09	0.13
0.5	0.05	0.06	0.09
0.6	0.03	0.04	0.06
0.7	0.02	0.03	0.04
0.8	0.01	0.02	0.02
0.9	0.01	0.01	0.01

■圖 7　研究假設之先驗機率的重要性

小結：當檢定強度或研究假設的先驗機率甚低的時候，$\alpha = 0.05$ 可能嚴重低估了僞陽性之反機率，也就是在 p 值檢定顯著的情況下，虛無假設 H_0 仍然

極有可能為真，而其為真的條件機率可能甚大於 α。此時如果我們拒絕虛無假設，便作出了錯誤的統計推論。

「摘櫻桃」問題

再來我們講到「摘櫻桃」問題，如同剛剛所提到，研究假設的先驗機率是如此重要，我們要如何去判定？要怎麼知道它是多少？我們必須要做文獻的分析、要建構我們的理論，在這種情況之下，會出現摘櫻桃的問題。這裡就是要呈現給大家看，譬如我們作 20 個統計檢定，從作第一個開始，本來有一個 model，但是 p 值不顯著，我們就改一下 model，加一個變數、減一個變數，或是把一個變數平方，或是把一個變數取 log，或者把樣本除去一些，增加一些，這樣慢慢去試驗，最後終於得到一個顯著的結果了！但這裡告訴你，做了 20 個這樣的檢定，我們以為每一個檢定的 type I error 控制在 0.05，可是 20 個裡面最少有一個顯著的或然率是多少？是 0.64（圖 8）。

多重 P 值檢定的「摘櫻桃」問題：
Cherry-Picking in Multiple Comparisons

- "Proper inference requires full reporting and transparency."
- If 20 tests are performed whose null hypotheses are all true, what is the probability that at least one test will be statistically significant?
- If the tests are independent, the probability is:
- $\Pr(\text{at least one test is significant})$
  $= 1 - \Pr(\text{none of the tests are significant})$
  $= 1 - (1 - .05)^{20} = .64$

■圖 8　多重 p 值檢定的「摘櫻桃」問題

為了讓大家能夠進一步瞭解這個問題，再給大家看一部電影，這部電影是「班傑明的奇幻旅程」。（電影「班傑明的奇幻旅程」短片連結：https://www.youtube.com/watch?v=mTDs0lvFuMc）

讓大家看這部電影，我們可以注意到，這部電影所講的，跟上一部「玉蘭花」很類似，也在討論是這樣發生車禍到底是 by accident 還是 by design。它的議論應該是：這種車禍的發生，其實有一連串的因果鏈，只要這因果鏈其中有一個環節稍微不一樣、或是沒有發生的話，可能車禍就不會發生。因此它的敘述者暗示說其實是 by design，而不是 by accident。然而現在要跟大家說明，這個結論是錯的。電影要說明這是 by design 而不是 by accident 的話，是完全錯誤的。爲什麼？大家只要想想看，我們政大門前有條交通繁忙的馬路，你一邊跳舞一邊過街，看會不會被車撞上，不是極有可能會嗎？爲什麼？因爲說車禍是 by accident，它是說被某一輛特定車子撞到的機率很低，譬如是 0.05，可是如果有 20 輛車子經過的話，被其中最少一輛撞到的機率就會很大，剛才已經算給各位看，所以電影是錯誤的。

類似這種問題，其實我們日常生活中所在多有。再以大樂透爲例：你買了一注大樂透，你中頭獎的機率是 1/13,980,000。如果你自己中獎，你也許會說這是命運，不是機率，因爲中獎的機率近乎 0。但全台灣賣了 5,000,000 注的大樂透，最少有一注中頭獎的機率其實是 0.30。你不能舉出有人中獎的事實就否定大樂透開獎的隨機機制。

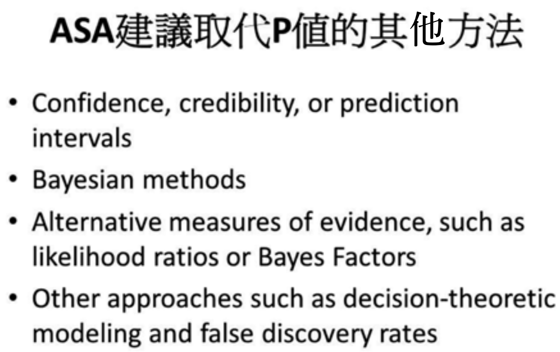
這就是 cherry-picking，只抓住發生的事件，就來說因爲有這麼多因果鏈，如果稍微有一點不一樣，這種事情就不會發生，這是錯誤的，因爲它有很多其他的可能性同時存在。現在在統計學裡面，很多人很不在意這個問題，甚至主張這種問題不存在，而其實它可能比 p 值的誤用還要嚴重。這種問題叫做叫多重假說檢定（multiple hypothesis test）、多重比較（multiple comparison），我有同事對這種問題的反應十分強烈，主張所有的研究都必須要事先登記，什麼叫做事先登記？並非申請研究經費、寫一個研究計畫這麼簡單，所謂事先登記（pre-registration）的觀念，就是在做任何研究之前，研究者必須要把研究計畫 post 在網站上，而且 post 上之後就不能改，現在其實已經有很多這種網站存在，將來研究者發表文章，如果跟預先登記的研究設計不一樣，其他人就可以對你發表的結果提出質疑。

小結：在多重假說檢定的情況下，即使 H_0 爲眞，「至少有一 p 值檢定顯著」的機率常會甚大於單一 p 值檢定的顯著水平 α。以「摘櫻桃」的方式只報

告顯著的檢定結果常會導致錯誤的統計推論。

結語

圖 9 是 ASA 建議取代 p 值的其他途徑，在此沒有時間細講，大致上是要用其他方法，比如貝式統計學。這邊提到的很多方法都跟貝式統計學有關係。我們現場有貝式統計學的專家，他們懂得怎麼用貝式統計學來分析資料。但對於還沒有學到貝式統計學的朋友，這邊 ASA 特別提到的 confidence intervals——信心區間–是傳統統計學的方法。ASA 似乎認為使用信心區間比使用 p 值檢定要來得好。但是信心區間其實是連續性的 p 值檢定，如果只是看看虛無假設的理論值有沒有在信心區間之內，則檢定的結果跟 p 值檢定是一樣的。但如果把信心區間畫出來，至少有一個好處，它會清楚呈現出效應的大小，讓你不但能看出檢定結果的統計顯著性（statistical significance），也能看出估計值的實質顯著性或重要性（substantive significance）。我們使用信心區間，總比只用一顆星兩顆星來標明統計顯著性要好。



■圖 9　ASA 建議取代 p 值的其他方法

如果一定要用幾顆星的話，大家就不要再用 $\alpha = 0.10$ 了；$p < 0.10$ 就不要再加星星了。我知道 AJPS——*American Journal of Political Science*——已經不接受 $\alpha = 0.10$ 這個顯著水準的統計檢定了；不管是單尾檢定或是雙尾檢定，用

$\alpha = 0.10$ 已經不被接受了。0.05 還可以，最好能用 0.01，審稿人對你較難有所批評。

但是最重要的，如果我們不得不用傳統的統計方法，我們必須要增強我們的理論論述和脈絡描述，因爲增強理論論述和脈絡描述，即會增強研究假設的先驗機率。當研究假設的先驗機率比較高時，其後驗機率–僞陽性的反機率–就會比較低。這好比你健康檢查某種疾病的篩檢出現陽性時，好的醫生會從你的性別、年齡、生活習慣、飲食作息、家庭病史、乃至於居住環境等脈絡來判斷你是否有充分的病因，以之來詮釋篩檢的陽性結果。這其實就是貝氏更新的道理。

我讀這些文獻後的想法是：統計學很快就會有很重大的改變，傳統的作法、用 p 值來作統計檢定的作法，大概再過幾年，就不容易再存在。所以大家必須要應變，這也是我在回國來，希望能夠提醒大家注意的一個問題。

Q&A 時間

提問一：

林老師您好，謝謝您今天很精彩的演講，也很謝謝上禮拜六參加計畫時，您給我們的文章有很大的啓發與提升。今天聽了這個演講以後，我覺得我們對於 p-value 的使用可能要有心理準備，未來就算不是被全部淘汰，大部分也要被丟到另外一邊去。我在想的一個問題是，因爲老師提到使用 confidence intervals，我們在寫作時，有一個習慣是會比較傾向去解釋那些在 p-value 上顯著的變數，如果說未來使用 confidence intervals 的話，我們是不是應該在文章裡面，每一個變數都要去解釋它對 dependent variable 的重要性？或是說應該怎樣去作結果的討論以及處理？謝謝！

回答一：

我想你的自變數應該也有所謂的解釋變項與控制變項吧。我覺得如果控制變項不是麼重要的話，也許就不用太費勁去討論，就著重在解釋變項。解釋變項就是不管作傳統的統計顯著或不顯著，都要加以討論。不只是討論統

計的顯著性，更要討論實質的顯著性，而實質的顯著性或重要性是比較能從 confidence intervals 看出來的。其實 p 值的問題是兩面刃，說不定對我們也有好處，就是將來得到不顯著的結果，說不定都可以 publish，都可以呈現在你的論文裡面，而不用怕被人家說：明明就不顯著爲什麼還要報告。

提問二：

林老師您好，我是經濟系的學生，謝謝林老師今天很精彩的說明，但這邊至少有兩個點想跟林老師請教，以及跟大家分享。第一個就是如您剛才所說，我們在作實證研究的時候，不管是我們自己或是長期的訓練，或是目前的期刊的要求，關切的都比較是顯著的結果。所以過去在經濟學界也有對這方面的討論，談到爲什麼要去關切那些不顯著的結果。同樣的道理，那些不顯著的結果要被期刊接受的機會也是非常非常低。你唯一可以被接受的理由大概就是，我們看到這個人所作的東西，以後就不要再作了，大概就是樣子。我第一點要說的是，我們目前有這樣的困境。您剛提到一個很好的論點，未來也許大家會有一個共識，就是不顯著的結果反而是更重要的。我的第二點是一個問題：您剛剛提到，確實在醫學或自然科學部分，要去找到一些理論上的基礎，可能相對來講比較容易。在社會科學裡面，如果要去找到一些所謂的因果關係，或是比較扎實的理論，可能比較困難，因爲人的行爲無法像自然科學的實驗室般重複去作，且控制到所有條件都一樣。針對此部分，您剛認爲要加強理論的論述，好讓 prior 來的比較 solid 一點，就社會科學部分不知道有沒有更好的一些方法，或至少不會差自然科學太多？這部分確實對我們社會科學的人來講比較困擾一點。

回答二：

我先從第二個問題來回答。我不敢說整個社會科學啦，但在政治學界大概很多人會跟你說：你可能要用賽局理論。美國政治學在過去十幾年來有一個概念叫作 EITM——Empirical Implications for Theoretical Models。名稱有點奇怪，但它的用意是把統計分析跟理論結合，講 EITM 的人特別強調的就是形式理論，特別是賽局理論。就是作一些對人性的基本假設，然後用賽局理論的數

學分法去 deduce，用邏輯去導出一些結果出來，然後再把這些結果用統計方法加以檢定。這在政治學過去十幾年來，已經變成一個很普及的概念。這有它的好處，就是在形式理論部分，只要基本假設大家能接受，它的邏輯都是沒有爭議的。嚴格來講，形式理論只要大家接受你的假設和邏輯推演，就要接受你的結果，用統計來檢定結果是多餘的。但是我們知道，比如假設行爲者是理性的，然而眞實的人不一定理性，所以經驗檢定還是重要的。EITM 用形式理論來增強理論的先驗機率，我想這是很不錯的。

你前面第一點提到關於不顯著的結果，當然我也不是說將來學術期刊會大量接受不顯著的檢定結果，我想也不至於，可能只是要求你把這些不顯著的結果都 post 在網頁上。然而對於教授升等，這些作品算不算也不一定。但是我想某種程度上這是合理的預期，一旦不需要使用幾顆星的話，不顯著的結果也可以放進文章裡去。期刊會衡量從整篇文章的研究設計、立論、方法、和結果，來決定到底能不能發表，而不會斤斤計較是一顆星、兩顆星，還是沒星星。所以我對這點倒是有點樂觀。其實，現在已經有很多期刊採取「預約接受刊登」（pre-acceptance）的編輯政策，也就是審查你的研究計畫就可以決定要刊登你計畫執行後的完稿，條件是不論經驗資料支持不支持你的研究假設，完稿都不得改變當初的研究設計，包括 model specification。這就是說不顯著的結果也要刊登了。

其實可以跟大家預告一下，8 月 4 日在中央研究院政治學研究所，爲了慶祝所慶，有一個學術討論會。討論會的主題是「甚麼是研究發現」？引言人有朱雲漢、吳玉山兩位院士跟我三個人。我的任務就是報告 p-value 的問題。傳統來講，統計上顯著的結果才叫做 findings，不顯著的結果是 non-findings，但是這觀念可能要有所改變了。這等到 8 月 4 日再專門來講。

提問三：

謝謝林老師很深入淺出的演講，之前在上統計課的時候，雖然有講到 p-value 的問題，但每次在上大學部課程時，我常常都沒辦法把這一塊講得這麼清楚。在我還是研究生的時候，我們就有很多這方面的討論，而這幾年這問題特別地被突顯，我認爲很大的原因，大概是電腦技術愈來愈好、作 testing 的

困擾已經愈來愈少；另一方面，如果你相信 Bayesian 的話，你應該相信所有的 parameters 都該是 probability term，而不是 deterministic term，說它是顯著還是不顯著。我也有一個問題想請教林老師，您如今在基礎統計的教學裡面，對 p-value 是用傳統 frequentist 的講法，還是像現在等於把它推翻？因爲我常有這樣的困擾，就是在初級的課用 frequentist 的方式講，然後到了進階的課，再拿 Bayesian 的 approach 去推翻自己原本以前講的。我不知道林老師您目前在授課時，是用什麼樣的方式？特別是針對 frequentist 的邏輯。

回答三：

我想你對 p 值問題的瞭解應該比我更早。我是這幾年來才慢慢地逐步瞭解這個問題。在教學上要採取立即的改變，其實很不容易，我完全瞭解。我們有一個同事後來就在抱怨，ASA 爲什麼要發表這個東西？他說現在所有的 journal articles，還有教材、教科書，全部——至少 90 幾 %——都是傳統的統計學，你怎麼來教大學生新的東西？所以這是很困難的。今天我在這裡演講，如果有一點點是我自己觀察來的結果，而不是完全從文獻上得到的，我想是關於 prior——HA 的 prior——怎樣去影響到僞陽性的反機率，這我覺得很重要。我目前教學仍是會用傳統方法，畢竟要把一本教科書重新編輯、作講義，是很大的工程。此外，我自己跟你不一樣，我是 frequentist，你來教 Bayesian 比我容易多了。我以前會放電影，跟學生講 p 值是什麼。我現在也放電影，跟學生講 p 值有什麼問題，讓他們瞭解。然後我會對他們說，在還沒學習貝式統計學之前，要比較強調 prior。也就是你用傳統的統計方法作研究，如果研究假設沒有很高的 prior 的話，也許你就不要作了。

提問三（接續）：

我只是有時候會有點精神錯亂，之前跟學生講過的東西，在比較進階的課程時就要把它推翻掉。

回答三（接續）：

在座如果有老師教統計學，請你不要說：林老師今天講的就代表我上課講

的都錯了。學生也不要說我上課學的都錯了。不是這麼一回事，這不是我的用意。因爲 p 值本身它並沒有錯，錯的是大家對它的誤解誤用。至於傳統的教學方法要怎麼改，我們要慢慢試，但是我們要瞭解這個問題的存在。我自己到最近教學還是用傳統方法，如果今天請我的學生來聽我演講，他們會說：老師你以前教的都錯了。但事實上，不只是我們教書的，有多少科學、商業或政策上的決定，都是奠基於 p 值檢定的結果之上，我們能說他們都錯了嗎？我想不能說他們都是錯的，可是我們要改變。

提問四：

林老師好，我是理學院資科系的老師。非常謝謝林老師，很高興今天上老師的課。關於剛剛幾位老師的討論，我覺得在我們資科系，很多人的直覺，一個方法要嘛是對、要嘛是錯。你們搞機率的卻是：它可能 80% 對、20% 錯。我覺得應該講清楚的是，就 prior 來講，只要 prior 夠強，過去 p-value 的方法大概是對的。這應該有 range，大部分問題，只要 prior 在 range 裡面，或許 p-value 的方法是相當可靠的。我不會推翻過去的教學方法，說一切都是錯的，其實沒有麼嚴重。在大部分的問題裡面，過去的方法也許是可用的，只是今天我們面對一些方法，單獨的 p-value 並不是麼可靠。也就是一個漸進式的改變，這樣我們不會打自己嘴巴。

回答四：

對，我完全同意。這就是爲什麼我做了這三個圖表，可以看到雖然影響僞陽性反機率的因素包括 prior 和 power，但其實主要是 prior。即使 power 低到 0.50，只要 prior 也有 0.50，僞陽性的反機率也不過是 0.09。如果你願意用 0.10 的顯著水準，0.09 還是顯著的！要給一個可接受的 range，我覺得 prior 大於 0.50 的話，其實都還好。最怕的就是 prior 很低很低，像 ESP 這種研究假設。這也是爲什麼在 p-value 問題的討論上，那一篇知名心理學家對 ESP 作的研究會被拿出來討論，因爲它的 prior 幾乎是零。但是這只能夠很粗略的估計。

提問五：

老師，這邊有一個小問題是：假設現在有十篇從舊到新的文章，它們的先驗機率都不太一樣，我如果要寫一篇文章，我要用最新一篇的先驗嗎？還是由自己發展出來、自己認定？

回答五：

當然你說先驗機率不太一樣，它為什麼會不一樣？是因為理論根本不一樣嗎？還是說因為時間的關係，大家有愈來愈多的研究發表，先驗機率就會逐步改變？如果已經有一個文獻，通常是建議你要作後設研究，叫 meta-analysis，就是把過去發表的文章統一起來作一個研究。但坦白說我個人也沒有作過這種 meta-analysis，可能可以在這方面的文獻去看一下。Eric，你可以就 meta-analysis 這點再作補充？

俞振華：嘗試把各種不同的 model 的係數，最後統整，變成有點類似老師您剛提的，試很多的 model 的 specification，然後組成一個結果。

林澤民：對，我讀的這些 p-value 的文獻裡面，其實有些文章就是作 meta-analysis。

提問六：

我有兩個關於寫作的問題。因為從老師的演講得到非常多心得，其中有個問題是，如果能強調理論先驗機率的強度，老師剛有提到用 EITM 看能不能夠結合形式理論的一些邏輯去增強強度，此外，我在思考是否有可能，至少就我自己在寫作時，會提出一些案例，然後再稍微說明，我有些案例，當然這些案例可證的是少數，因為全世界有一百多個國家，我們只有一兩個案例而已，說服力有限，但多多少少還是有些用處。我在想這樣作是否 OK？這是為了提升理論先驗機率的說服力，而提出一些案例來作討論。第二，剛剛老師提到有關 non-findings，這些發現，相信以後應該愈來愈多人至少在文中會提到，可能一段、或幾句話。就老師的想法來說，要提是要怎麼提？是跟目前為止像跟大家講的一樣，要提的話就只能說，結果顯示並不是 statistically significant，就這樣子很平鋪直敘的描述？還是要稍微把重點放在跟理論的連結，即便結果沒

有很顯著，但也不代表我的理論是錯的。我不曉得能不能這樣講，也許不行，因爲太武斷。只是不曉得未來大家在強調沒有統計顯著水準的結果時，是要怎麼表達？是要平鋪直敘地講，還是要有些焦點？有些要強調、有些不一樣？

回答六：

我想先講第二個問題，而其實這在 Bayesian 根本就不是問題，Bayesian 就把 posterior distributions 畫出來就好，你根本也不需要去提是否顯著，因爲「顯著」的概念本來就是 frequentist 的概念，它不是 Bayesian 的概念。所以要是你看過一些 Bayesian 的文章，你會看到它畫很多圖，每個圖都很小，一小格就一個圖，然後圖就畫上 posterior distributions，甚至連 credible intervals 也不一定要畫出。

俞振華：但是爲了要跟 frequentist 對話，現在還是會有 95% 的 credible intervals。

林澤民：對，不過需要 95% 的嗎？因爲我最近寫一篇文章，合作者說 68% 就可以。所以我想可能就不需要去談什麼顯著不顯著，你就把圖畫出來就好。你若不是 Bayesian，就用 confidence intervals，然後你去畫圖，每一個變數的係數你就把 confidence intervals 畫出來。至於 0 有沒有在 confidence intervals 裡面，我想不必然是唯一的重要標準。當然就實際情況來說，仍要看你的 reviewers 有沒有接受你的結果。我必須要強調，在網路上你還是可以找到一些文章，它們要替 p-value 辯護。要是碰到這樣的評論者，可能就必須要小心。你第一個問題是說，提出實質案例而不一定是理論，我覺得也可以，我個人會接受，因爲所謂文獻，除了理論之外，還有這種實質的知識、地方性的知識。我個人認爲這些知識可以幫助我們加強 prior，特別是當這些案例能夠增加我們瞭解自己研究假設的脈絡時。ASA 的聲明特別提到脈絡（context）的重要性，我剛剛也有提到醫生詮釋陽性反應時，通常要參考病人所處的脈絡。但是我必須要說，我今天特別強調 prior 的重要性，我不知道在座是否有其他學者可以肯定我這一點，我覺得我個人強調 prior，可能與文獻上的這些在講 p-value 的危險性的 articles 相較時，我強調的可能比較多一點。我不能保證所有的統計學者都會同意我的看法，所以要是碰到我來評審你的文章就好了。

但是我希望我講的還是有點說服力吧？要是你研究假設的 prior 夠強，可能 p-value 的問題就不是這麼大。

提問七：

聽了很多同仁的問題，還有老師的回答以後，我這邊另外的問題是，因爲在一開始，老師提到一個期刊——Basic and Applied Social Psychology，也講了 ASA 在今年提出的聲明，我想問，ASA 它的官方期刊——JASA，是否已經有接受，或是應該說拒絕這種只報 p-value 的文章？還是說他們政策現在是做一個調整，同時都接受兩種？

回答七：

很抱歉，JASA 的文章我不是經常在看，我不能回答你的問題。但是我剛剛已經講了，BASP 在他們政策制定之後，ASA 有一個回應，不是那 official statement，是在發表 official statement 之前的一個回應。那個回應只說 ASA 正在籌擬一個 official statement。而最後這 official statement 其實跟 BASP 的決定是不一樣的。因爲 ASA 的 official statement，第一點在說明 p-value 是什麼，它並沒有說 p-value 錯誤。它只是把 p-value 的正確意義講出來。換句話說，只要是使用正確的意義，p-value 並沒有問題，只是不要去誤用它。不要只是著重在統計顯著性，因爲 model 對錯的機率跟 p-value 不一樣。要使用 p-value 作檢定，要把它跟 α 來做比較，所以問題不只是 p-value，而是 α。界定了 α 之後，才知道結果是不是顯著。當得到一個顯著的結果以後，必須再來衡量偽陽性反機率的問題，也就是 model 後設機率的問題，這就不是 p-value 可以告訴你的。