數學傳播　41卷2期, pp. 64-69

如何設計滿足辛普森悖論的例子

連威翔

一、前言

筆者在數學網站昌爸工作坊 ([1]) 上, 發現一個有趣的統計例子, 此例子是兩個班級的考生於考試後錄取的結果, 其中除了分別列出兩班男、女生的報考人數與錄取率外, 也列出了總錄取率, 如下表:

表1

	報考人數	錄取率	總錄取率
甲班	男 10 人	$x_1 = 7/10$	$p_1 = \dfrac{22}{40}$
	女 30 人	$y_1 = 15/30$	
乙班	男 30 人	$x_2 = 19/30$	$p_2 = \dfrac{23}{40}$
	女 10 人	$y_2 = 4/10$	

(爲求簡化, 上表中未列出錄取人數, 讀者可簡單求得各錄取 7, 15, 19, 4 人, 底下各表同), 注意兩班男、女生分組錄取率滿足 $x_1 > x_2$, $y_1 > y_2$, 但總錄取率卻有 $p_1 < p_2$。換言之, 甲、乙兩個母群體, 各分成兩個對應的小集團, 局部對應做統計比較, 甲勝過乙, 但是整個合起來比較反而是乙勝過甲, 這種違反直觀的結果叫做辛普森悖論。在本文中, 筆者想研究此類例子是如何設計出來的。

二、探索

一般來說, 仿照上面表 1, 其實我們想設計出下面的例子:

表2

	報考人數	錄取率	總錄取率
甲班	男 n_1 人	x_1	$p_1 = \dfrac{n_1x_1 + m_1y_1}{n_1 + m_1}$
	女 m_1 人	y_1	
乙班	男 n_2 人	x_2	$p_2 = \dfrac{n_2x_2 + m_2y_2}{n_2 + m_2}$
	女 m_2 人	y_2	

其中正整數 n_1, m_1, n_2, m_2 數目待定, 而錄取率需先給定, 仿照表 1, 假設也滿足 $x_1 > x_2$, $y_1 > y_2$ 且 $x_1 > y_1$, $x_2 > y_2$。我們希望可找到適當的 n_1, m_1, n_2, m_2 之值, 使得

$$\frac{n_1x_1 + m_1y_1}{n_1 + m_1} = p_1 < p_2 = \frac{n_2x_2 + m_2y_2}{n_2 + m_2} \tag{1}$$

如同表 1 的例子, 我們假設表 2 兩班總人數相同, 都是 S 人, 即

$$n_1 + m_1 = n_2 + m_2 = S$$

此時, 注意原本分組錄取率條件爲 $x_1 > y_1$, $x_2 > y_2$, 因此錄取率 p_1 滿足

$$\begin{aligned} y_1 = \frac{0x_1 + Sy_1}{S} < \frac{n_1x_1 + m_1y_1}{n_1 + m_1} < \frac{Sx_1 + 0y_1}{S} = x_1 \\ \Rightarrow\ y_1 < p_1 < x_1 \end{aligned} \tag{2}$$

同理, 錄取率 p_2 也會滿足

$$y_2 < p_2 < x_2 \tag{3}$$

此時筆者發現表 1 滿足 $p_1 < p_2$ 的關鍵, 在於有 $x_2 > y_1$ 的條件, 爲什麼呢? 增加 $x_2 > y_1$ 的條件後, 加上原條件 $x_1 > x_2$, $y_1 > y_2$ 就有 $x_1 > x_2 > y_1 > y_2$, 再由 (2), (3) 式所得 p_1, p_2 的範圍, 可知六個錄取率的初步關係將如下圖:

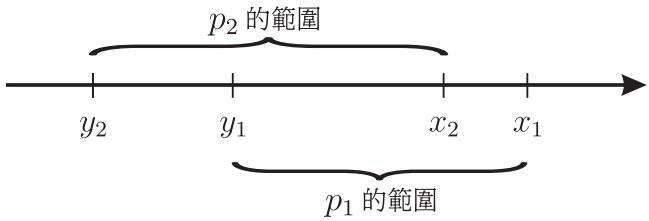

圖 1

注意 p_1, p_2 不能落在所在範圍的端點上, 雖然如此, 由圖 1 看來, 只要 p_1 夠靠近 y_1 且 p_2 夠靠近 x_2, 是有可能發生 $p_1 < p_2$ 的。換一個說法, 正是因爲 $x_2 > y_1$ 的條件, 使圖 1 中 (y_2, x_2) 和 (y_1, x_1) 兩區間有所重疊, 導致 $p_1 < p_2$ 有可能發生。

除了從圖 1 來看, 我們也從式子來看。因爲 $x_1 > y_1$, $x_2 > y_2$, 表 2 中錄取率 p_1, p_2 分

別滿足

$$p_1 = \frac{n_1x_1 + m_1y_1}{n_1 + m_1} = \frac{n_1x_1 + (S - n_1)y_1}{S} = y_1 + \frac{n_1(x_1 - y_1)}{S}$$
$$\Rightarrow\ p_1 - y_1 = \frac{n_1}{S}(x_1 - y_1) > 0 \tag{4}$$
$$p_2 = \frac{n_2x_2 + m_2y_2}{n_2 + m_2} = \frac{(S - m_2)x_2 + m_2y_2}{S} = x_2 + \frac{m_2(y_2 - x_2)}{S}$$
$$\Rightarrow\ x_2 - p_2 = \frac{m_2}{S}(x_2 - y_2) > 0 \tag{5}$$

在 (4) 式中的 $x_1 - y_1$ 與 (5) 式中的 $x_2 - y_2$ 兩者均爲定值, 因此只要使 $\frac{n_1}{S}$, $\frac{m_2}{S}$ 兩數之值任意小 (最簡單的取法就是取 $n_1 = m_2 = 1$ 且 S 儘量大), 就可使 p_1 任意靠近 y_1 (且 $p_1 > y_1$)、p_2 任意靠近 x_2 (且 $p_2 < x_2$), 此時因爲 $x_2 > y_1$, 由圖 1 知必能得到

$$y_1 < p_1 < p_2 < x_2$$

此結果可示意如下:

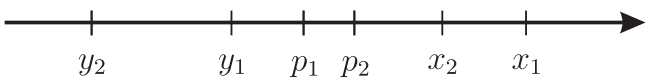

圖 2

眼尖的讀者, 看到上面 $p_1 = \frac{n_1x_1 + m_1y_1}{n_1 + m_1}$, $p_2 = \frac{n_2x_2 + m_2y_2}{n_2 + m_2}$ 兩式, 或許會想到這是數線上兩點間分別按照 $n_1 : m_1$ 和 $n_2 : m_2$ 的比例取分點坐標 p_1, p_2 的公式, 而圖 2 其實就是筆者配合表 1 的例子 $(n_1 = 10, m_1 = 30, n_2 = 30, m_2 = 10)$ 而畫, 其中取區間 (y_1, x_1) 的 3 個四等分點 $(n_1 : m_1 = 1 : 3)$ 中最左邊的點爲 p_1, 並取區間 (y_2, x_2) 的 3 個四等分點 $(n_2 : m_2 = 3 : 1)$ 中最右邊的點爲 p_2。

接下來的過程, 是爲了估計出兩班總人數 S 至少要多大, 才能得到像圖 2 那樣的結果。如果先給定表 2 中四個分組錄取率 x_1, x_2, y_1, y_2, 其中 $x_1 > x_2 > y_1 > y_2$, 若存在正整數 n_1, m_1, n_2, m_2 滿足 (1), 我們會有:

$$\frac{n_1x_1 + m_1y_1}{n_1 + m_1} < \frac{n_2x_2 + m_2y_2}{n_2 + m_2} \tag{6}$$
$$\Leftrightarrow \frac{n_1x_1 + (S - n_1)y_1}{S} < \frac{(S - m_2)x_2 + m_2y_2}{S}$$
$$\Leftrightarrow y_1 + \frac{n_1(x_1 - y_1)}{S} < x_2 - \frac{m_2(x_2 - y_2)}{S} \tag{7}$$

因爲已知 $x_1 > x_2 > y_1 > y_2$, 此時假設

$$x_1 - y_1 = \delta > 0,$$
$$x_2 - y_2 = \ell > 0,$$

將上述兩式代入 (7) 後可繼續推得:

$$\begin{aligned} & y_1 + \frac{n_1\delta}{S} < x_2 - \frac{m_2\ell}{S} \\ & \Leftrightarrow x_2 - y_1 > \frac{n_1\delta + m_2\ell}{S} \geq \frac{\delta + \ell}{S} \qquad (8) \\ & \Leftrightarrow S > \frac{\delta + \ell}{x_2 - y_1} = \frac{x_1 - y_1 + x_2 - y_2}{x_2 - y_1} = 1 + \frac{x_1 - y_2}{x_2 - y_1} \qquad (9) \end{aligned}$$

其中 (8) 利用到 $n_1, m_2 \geq 1$。上述 (9) 式就是表 2 關於 S 取值的必要條件, 不妨稱之爲 S 的取值條件式。若我們取表 1 中 x_1, x_2, y_1, y_2 之值做爲例子, 將其代入 (9) 可得

$$S > 1 + \frac{21 - 12}{19 - 15} = 3\frac{1}{4} \qquad (10)$$

從 (10) 式看來, 難道說兩個班級都只要有 4 個人就可以設計出表 1 那樣的例子? (表 1 的兩班總人數都是 40 人) 且讓我們沉住氣一下, 就先取 $S = 4$ 沒關係。注意在 (4), (5) 之後的討論, 我們知道要使 $\frac{n_1}{S}, \frac{m_2}{S}$ 兩數儘量小, 因此我們再取 $n_1 = m_2 = 1$, 此時表 2 變成底下的情形:

表 3

	報考人數	錄取率	總錄取率
甲班	男 1 人	$x_1 = 7/10$	$p_1 = \frac{22}{40}$
	女 3 人	$y_1 = 15/30$	
乙班	男 3 人	$x_2 = 19/30$	$p_2 = \frac{23}{40}$
	女 1 人	$y_2 = 4/10$	

注意表 3 中兩班總錄取率確實滿足 $p_1 < p_2$。但其中甲班男、女生與乙班男、乙生的「錄取人數」分別爲

$$\frac{7}{10}, \frac{15}{10}, \frac{19}{10}, \frac{4}{10}$$

此四數都不是正整數, 故表 3 的數據不合理。而剛好上面四個分數的分母皆爲 10, 因此只要將表 3 中"報考人數"欄位的四數全部乘以 10, 則上述四個錄取人數就變爲

$$7,\ 15,\ 19,\ 4$$

這就成爲了合理的數據, 而且所有錄取率都不變。因此透過"把人數乘以某個倍數"的方法, 我們可使原本不合理的數據合理化, 而透過此法, 表 3 也就變成了表 1。

三、模仿

除了研究表 1 的例子, 我們不妨自己練習造個例子, 如下表:

表4

	報考人數	錄取率	總錄取率
甲班	男 n_1 人	$x_1 = 17/18$	p_1
	女 m_1 人	$y_1 = 13/18$	
乙班	男 n_2 人	$x_2 = 16/18$	p_2
	女 m_2 人	$y_2 = 12/18$	

注意上表中我們先給定 4 個分組錄取率, 它們滿足 $x_1 > x_2 > y_1 > y_2$, 我們也希望有 $p_1 < p_2$。將表 4 中的 x_1, x_2, y_1, y_2 代入 S 的取值條件式 (9), 得到

$$S > 1 + \frac{17-12}{16-13} = 2\frac{2}{3}$$

因此暫取 $S = 3$, $n_1 = m_2 = 1$, 則 $m_1 = n_2 = 2$, 此時表 4 變爲

表5

	報考人數	錄取率	總錄取率
甲班	男 1 人	$x_1 = 17/18$	$p_1 = \frac{43}{54}$
	女 2 人	$y_1 = 13/18$	
乙班	男 2 人	$x_2 = 16/18$	$p_2 = \frac{44}{54}$
	女 1 人	$y_2 = 12/18$	

表 5 中總錄取率滿足 $p_1 < p_2$, 但其各組錄取人數不合理, 分別爲 $\frac{17}{18}, \frac{26}{18}, \frac{32}{18}, \frac{12}{18}$, 我們仿照之前的作法, 將表 5 中的報考人數乘以 18 倍, 即得到新的錄取人數爲 17, 26, 32, 12, 這樣就有了下表:

表6

	報考人數	錄取率	總錄取率
甲班	男 18 人	$x_1 = 17/18$	$p_1 = \frac{43}{54}$
	女 36 人	$y_1 = 13/18$	
乙班	男 36 人	$x_2 = 16/18$	$p_2 = \frac{44}{54}$
	女 18 人	$y_2 = 12/18$	

表 6 就是仿照表 1 造出的新例子, 同樣滿足 $x_1 > x_2$, $y_1 > y_2$, 但 $p_1 < p_2$。

四、結語

像表 1 與表 6 這樣滿足辛普森悖論的統計結果, 直觀上來說, 因兩表中甲班的女生人數佔大多數, 所以甲班總錄取率 p_1 會靠近該班女生的錄取率 y_1; 而兩表中乙班的男生人數佔大多數, 所以乙班總錄取率 p_2 會靠近該班男生的錄取率 x_2。而表 1 和表 6 一開始就設計讓 $y_1 < x_2$, 所以 (從圖 1) 也不難看出會有 $p_1 < p_2$ 的可能了。

經過上述的研究過程, 我們大約理解了這類統計例子的由來, 並且應該也能自己設計出其他的例子。之後若您有朋友看到此類例子而感到疑惑, 愛好數學的你, 或許可以有信心地告訴他們:「嘿! 朋友, 我知道那個例子是怎麼設計的, 讓我先來畫個圖 (圖 1), 再慢慢解釋給你聽, 好嗎?」

關於辛普森悖論更進一步的介紹, 有興趣的讀者可參考[2]。最後, 筆者想感謝昌爸在參考資料 [1] 上所提供的熱心協助, 也要感謝審稿人所提出的寶貴修改意見。

參考資料

1. 昌爸工作坊關於 Simpson's paradox 的介紹。
   http://www.mathland.idv.tw/fun/Simpson's_paradox.htm.
2. Wikipedia — Simpson's paradox.
   https://en.wikipedia.org/wiki/Simpson%27s_paradox.

—本文作者任職參當勞竹南民權中心—