民意調查

bee*

104.09.25

民意調查的統計理論架構[1]。

1. 前言

在高三上學期課程的第一章中，我們學了很多統計的理論，究竟這一章要告訴我們甚麼呢？本文用倒述式的方法，把整章說明一次，讓你可以很快地掌握其統計上的意義。

2. 民意調查

這一章的主要目的是希望我們透過「抽樣的方法」，去了解「母體的特性」。

何爲「抽樣」? 何爲「母體」? 何爲「母體的特性」? 舉個例子：

我們想知道「全台灣地區高中學生喜好數學的程度」。

1.「全台灣地區高中學生」，就是母體，顯然，母體中的高中生人數是相當的多。

2.「喜好數學的程度」就是母體的特性。

3. 爲何要「抽樣」，因爲母體的「樣本數實在是太多啦」? 只能抽取一部份的樣本來看看。

於是，我們就想辦法去拜訪 1000 位高中生，盡可能的把拜訪的範圍包含「城市和鄉村」，包含「各年級」，然後問調查的對象：你喜歡數學嗎？答案只有兩個選擇：喜歡或不喜歡，然後把答案記錄下來，然後統計出「喜歡數學的比例」。

最後，我們計算出一個「範圍」，稱爲「信賴區間」，然後用這一個區間，表示調查的結果。

---

*bee 美麗之家: http://www.beehome.idc.tw

[1]102 課綱的教材。

例如：說喜歡的有 550 人，於是：喜歡的比例爲 $\frac{550}{1000} = 0.55$，並計算一下：

$\sigma = \sqrt{\frac{0.55 \times 0.45}{1000}} \approx 0.0157$，因此，「全台灣地區高中學生喜好數學的程度」爲

$$(0.55 - 2 \times 0.0157, 0.55 + 2 \times 0.0157) = (0.5814, 0.5186)$$

在上面的過程中，抽樣後的統計比例爲 $\hat{p} = \frac{550}{1000} = 0.55$，標準差的計算公式爲

$$\sigma = \sqrt{\frac{p \times (1-p)}{n}}$$

然後用一個區間 $(\hat{p} - 2 \times \sigma, \hat{p} + 2 \times \sigma)$ 表示母體的特性 p 或落在這一個區間裡。

1. 特性比例：$\hat{p} = \frac{\text{贊成樣本數}}{\text{總樣本數}}$，如愛好度，支持度，滿意度。
2. 標準差：$\sigma = \sqrt{\frac{p \times (1-p)}{n}}$。
3. 信賴區間：$(\hat{p} - 2 \times \sigma, \hat{p} + 2 \times \sigma)$。這是專屬於 95% 信心水準的信賴區間。

要注意：母體的特性比例用 p 表示，樣本的特性比例則用 $\hat{p}$ 表示，原則上，p 會落在由 $\hat{p}$ 計算出來的信賴區間之間，這實在是非常棒的結果。

例題 **1.** 關於總統大選甲候選人的支持度調查，經成功調查 1100 位具有選舉權的公民得到 605 位支持，試寫出在 95% 信心水準的信賴區間。

計算 $\hat{p} = \frac{605}{1100} = 0.55$，$\sigma = \sqrt{\frac{0.55 \times 0.45}{1100}} = 0.015$，

因此，95% 信心水準的信賴區間爲 $(0.55 - 0.015 \times 2, 0.55 + 0.015 \times 2) = (0.52, 0.58)$。

■

在這一個部分，我們學會了計算信賴區間的方法，只要套公式就可以得到結論，但是，心中難免有很多疑惑：

1. 公式哪裡來？
2. 一千個左右的樣本數是固定的需求嗎？這樣本數不會太少嗎？
3. 爲何我們老是要求 95% 信心水準的信賴區間？
4. 95% 信心水準的信賴區間到底是啥意思？

3. 民意調查的意義

我們一一來回答上面的問題。

所謂的公式，其實是柏努利試驗算術平均數與標準差。也就是，所謂的抽樣，其實就是柏努利試驗，抽樣 1000 個樣本，就是做 1000 次的柏努利試驗。我們用符號 $B(n,p)$ 來表示柏努利試驗。其中 n 是做 n 次柏努利試驗，而 p 是柏努利試驗成功的機率。

因爲柏努利試驗是獨立試驗，所以 $B(n,p)$ 的期望值 $\mu_X = np$，而標準差爲

$$\sigma_X = \sqrt{np(1-p)}$$，其中 X 是成功次數的隨機變數

做民調時，我們把每一個人當作是支持度爲 p 的柏努利試驗，然後根據柏努利試驗的機率分布，我們得到的成功次數應該和 $B(n,p)$ 的期望值 $\mu_X = np$ 的差距不會太遠，如果做 1000 次柏努利試驗，那麼，這差距差不多就在 2 個標準差之間，這眞是一個很不錯的結論，如圖 1 所示：

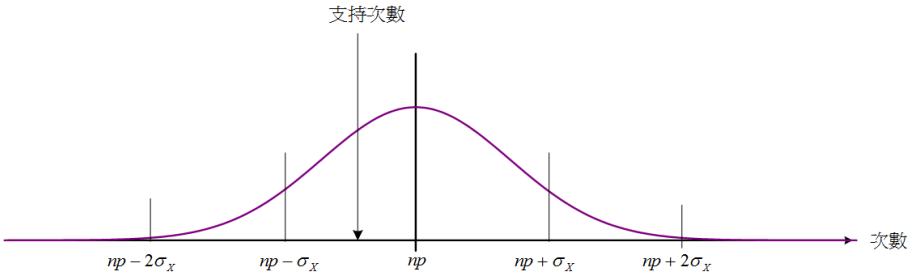

圖 1: 次數型的民調情形

如果我們把焦點從「次數」變成「比例」，那麼圖 1 會變成圖 2：

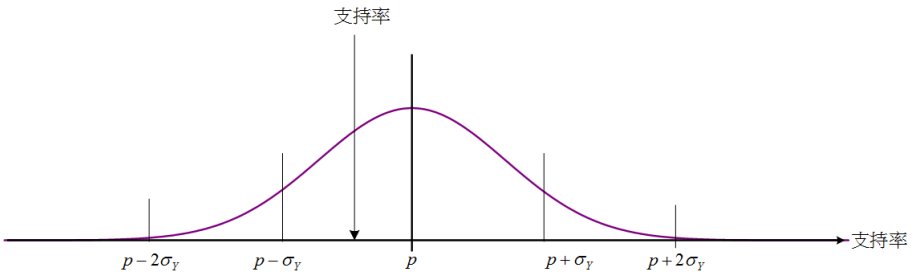

圖 2: 比率型的民調情形

如果研究的是比例，那麼隨機變數 X 就會變成 $Y = \frac{X}{n}$，也就是 Y 是「比例」的隨機變數。因此還是柏努利試驗，眞正母體的支持度 (特性比例) 爲 p，p 是我們不知道的數，但是，當我們做 1000 次柏努利試驗時，「抽樣的支持度 $\hat{p}$」就會和眞正的母體支持度的差距不會超過 2 個標準差，這時候的標準差是 Y 的標準差 (即 $\sigma_Y = \frac{\sigma_X}{n} = \sqrt{\frac{p(1-p)}{n}}$)。

本來母體的支持度 p 是主角，但是因爲我們不知道 p 的值，所以只好把 $\hat{p}$ 當主角，並因爲它們兩數的差距不會超過兩個標準差，所以 $|p-\hat{p}|<2\sigma_Y$，即 p 在區間 $(\hat{p}-2\sigma_Y, \hat{p}+2\sigma_Y)$ 內。由圖 2 我們知道，從 $p-2\sigma_Y$ 到 $p+2\sigma_Y$ 之間，幾乎涵蓋了所有可能的情形，所以，還沒有做民調之前，我們就可以知道：調查後所得的區間「 應該會涵蓋眞正的 p 值 」。

稍微理解了民調的道理了嗎？

1. 民調的公式就是柏努利試驗 $B(n,p)$ 的公式。
2. 95% 信心水準的信賴區間是指兩個標準差的半徑，此時區間的長度是 4 的標準差。由圖可知：2 個標準差的範圍很足夠了，令人很有信心。
3. 事實上，利用公式計算出來 95% 信心水準的信賴區間，大約每做 100 次民意調查，有 95 次會抓到母體眞正的支持度。不過，到底有沒有抓到，我們其實不知道，而「 事後 」也不能用機率表示，因此，只好用「 95% 信心水準 」這樣的名詞。

4. 再深入探討

由上面的討論可知：民意調查是建立在 n 次柏努利試驗上，但是，我們怎能計算 n 次柏努利試驗的機率分布呢？

很有意思的是：n 次柏努利試驗的機率分布可以用「 常態機率分布 」加以模擬。

常態機率分布是一個連續型的機率分布，其機率函數爲 $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu_x)^2}{2\sigma_x^2}}$，這是一個對稱於直線 $x=\mu_x$ 的圖形，因爲是「 指數的倒數型態 」，所以其圖形與 x 軸漸近，它的函數圖形非常漂亮，透過積分，我們可以得到函數以下與 x 軸圍成的面積爲 1，且透過積分可以知道：

1. $P(\mu_x-\sigma<x<\mu_x+\sigma)=\displaystyle\int_{\mu_x-\sigma}^{\mu_x+\sigma} f(x)dx\approx 0.68$；
2. $P(\mu_x-2\sigma<x<\mu_x+2\sigma)=\displaystyle\int_{\mu_x-2\sigma}^{\mu_x+2\sigma} f(x)dx\approx 0.95$；
3. $P(\mu_x-3\sigma<x<\mu_x+3\sigma)=\displaystyle\int_{\mu_x-3\sigma}^{\mu_x+3\sigma} f(x)dx\approx 0.997$。

如圖 3 所示：

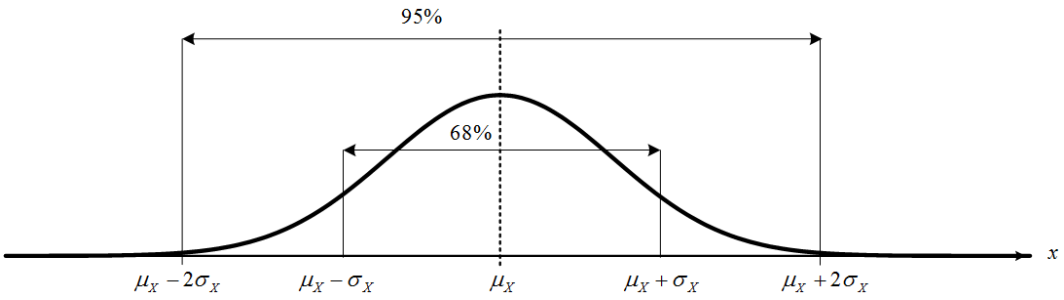

圖 3: 常態機率分布圖

這一個圖眞的非常漂亮，由圖 3 可以了解爲何常態分布曲線函數長成

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu_x)^2}{2\sigma_x^2}}$$

最後我們討論一件事：爲何老是抽樣 1000 個左右的樣本。

回到例題 1，關於這次的民意調查，報社的報導是：

本次調查：在台灣地區全體具有投票權的所有公民中，成功抽樣 1100 位公民，在 95% 的信心水準下，其中有 55% 的公民支持候選人甲，正負誤差爲 3%。

看看上面的例子，你應該很滿意這些數據，其中正負誤差爲 3%，讓我們覺得很安心，同時也可以確定候選人甲應該是會當選的。現在想想看：正負誤差 3% 是啥意思？

它是 2 個標準差的意思，而標準差的計算公式是 $\sqrt{\frac{p(1-p)}{n}}$，你可以發現標準差的大小是被 p 和 n 所決定的。因爲 p 是一個小於 1 的數，所以 $\sqrt{p(1-p)} < 0.5$，這樣子，我們就可以用 n 來控制標準差的大小。

於是在感覺上誤差不要超過 3% 的情形下，我們可得

$$\sqrt{\frac{p(1-p)}{n}} < \frac{0.5}{\sqrt{n}} < 0.015，$$

進而計算解得

$$n > 1111。$$

也因此，民意調查總是抽樣 1000 個左右的樣本數，而誤差通常在 2.8% 到 3.2% 之間。

5. 結語

從結論回來看理論根據，也是一件很有趣的事情。雖然中間有許多道理我沒有講得很清楚，但是，你應該可以理解課堂上的內容。

了解內容比套公式解數學問題其實更有趣的多，是嗎？