淺談自由度

(樣本標準差公式中的分母為什麼要採用 $n-1$)

江振東／政治大學統計系

我們都知道，在母體平均數 μ 已知的情形下我們可以利用 $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2$ 來估計母體變異數。但是在母體平均數 μ 未知的情形下，我們則改用 $\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 來作估計。由於 μ 未知，因此直觀上我們可以 $\bar{Y}$ 來作取代，但是分母為什麼也要調整為 $n-1$ 呢？由於 $\sum_{i=1}^{n}(Y_i-\mu)^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2+n(\bar{Y}-\mu)^2$ ，因此 $\sum_{i=1}^{n}(Y_i-\mu)^2\geq\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 。如此一來，我們就可以發現如果 $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2$ 是母體變異數的一個好的估計量的話，那麼 $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 顯然就會有低估的可能。如果改用 $\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 來作為估計量，低估的現象應該可以獲得改善。但是為什麼是 $n-1$ ，而不是 $n-2$ ，甚至 $n-3$ 呢？這就牽涉到自由度的問題。

幾乎所有初等統計學的書本都會告訴讀者，統計量 $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 的自由度為 $n-1$ ，但是究竟什麼是自由度，而統計量 $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 的自由度又為什麼是 $n-1$ ，而不是 n，常常並沒有清楚的交代。實際上，如果假定 $y_1,...,y_n$ 是由同一母體產生的一組隨機樣本，我們可以發現 $\sum_{i=1}^{n}(y_i-\mu)^2$ 其實是向量 $\mathbf{y}_n^{\mu}=(y_1-\mu,...,y_n-\mu)'$ 的長度平方，而 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 則是向量 $\mathbf{y}_n^{*}=(y_1-\bar{y},...,y_n-\bar{y})'$ 的長度平方。然而這兩個向量是有些不同的。 $\mathbf{y}_n^{\mu}$ 很容易可以看出是 R^n 中的一個向量，而 $\mathbf{y}_n^{*}$ 雖然也是 R^n 中的一個向量，**實質**上卻是侷限在一個 $n-1$ 度的子空間裡。也就是因為這項差異，導致於我們需要將統計量 $\sum_{i=1}^{n}(Y_i-\mu)^2$ 除以 n，但是就統計量 $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 而言，卻改為除以 $n-1$ 的主要原因。以下我們將藉由向量代數的觀點，來說明自由度的概念。令

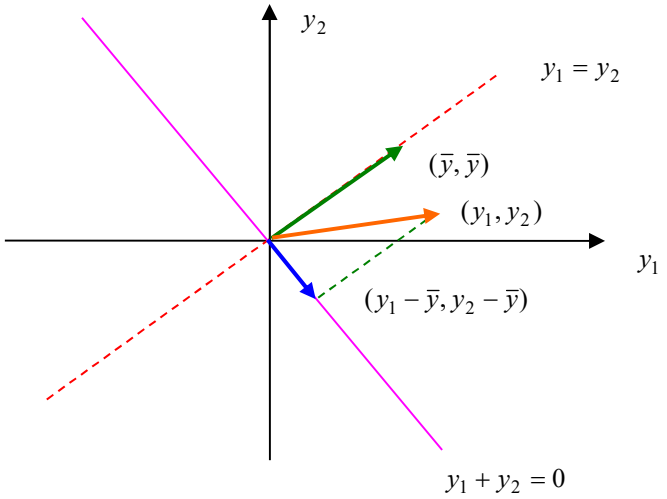
$$E_n=\left\{\mathbf{e}_n=(e_1,...,e_n)'\Big|\sum_{i=1}^{n}e_i=0\right\}\ ,$$

亦即 E_n 是所有滿足 $\sum_{i=1}^{n}e_i=0$ 這個 $n-1$ 度空間方程式的 $\mathbf{e}_n$ 所構成的子空間。我們可以注意到 $\mathbf{y}_n^{*}$ 其實就是其中的一員，而我們也將會發現到自由度指的就是子空間 E_n 的維

度。

在後續的討論中，我們將分成 $n=2$ 、 $n=3$ 與一般化的情形來作探討。 $n=2$ 的情形，因為可以繪圖，而且數學運算直接了當，因此對高中學生而言，應該不至於造成困擾。而 $n=3$ 的情形，雖然困難度稍微增加，不過概念完全是相通的，因此我們希望透過老師們的解說，學生們也可以了解其中的概念。至於一般化的情形，我們的對象則是老師，由於高中數學教師們在大學時期，一定學過線性代數，因此希望藉由這些說明，能夠讓老師們更清楚了解自由度的概念，從而協助學生們了解除以 $n-1$ 的原因。

我們就先從 $n=2$ 的簡單情形談起。儘管在實際生活裡，我們不可能僅僅選取兩個樣本，然而藉由 $n=2$ 這個特殊情形的探討與解釋，我們可以很容易的對任意 $n>2$ 的情形作衍生推論。



由上圖中，我們可以發現就任意的 y_1,y_2 而言，向量 $(y_1,y_2)'$ 一定可以分解成 $(\bar{y},\bar{y})'$ 和 $(y_1-\bar{y},y_2-\bar{y})'$ 這兩個正交向量的和。考慮 $\sum_{i=1}^{2}(y_i-\bar{y})^2$ ，這是 $\mathbf{y}_2^*=(y_1-\bar{y},\,y_2-\bar{y})'$ 這個向量的長度平方。由於

$$y_1-\bar{y}=y_1-\frac{y_1+y_2}{2}=\frac{y_1-y_2}{2}$$

$$y_2-\bar{y}=y_2-\frac{y_1+y_2}{2}=\frac{y_2-y_1}{2}=-\frac{y_1-y_2}{2}=-(y_1-\bar{y})$$

因此在 y_1,y_2 給定的情況下， $y_i-\bar{y}$ 其實都是 $\frac{y_1-y_2}{2}$ 的形式，具有相同的絕對值。這也就是說如果 $y_1-\bar{y}$ 或 $y_2-\bar{y}$ 的其中任意一個值一旦給定，那麼另一個值也就被決定了，兩者之間只是一個正負號的差異而已。同時這也就意謂著 $(y_1-\bar{y},y_2-\bar{y})$ 這個點

一定在 $y_1 + y_2 = 0$ 這條直線上，這件事實我們也可以由上圖中觀察得到。因此 $\sum_{i=1}^{2}(y_i - \bar{y})^2$ 雖然是兩個項的平方和，但是由於

$$(y_1 - \bar{y})^2 = (y_2 - \bar{y})^2$$

因此在作計算求和時，我們只需**自由**選擇其中一項來作計算即可，並不需要重複再計算另外一項，也就是說其中一項的平方和可以由另外一項平方和完全決定；再者由於 $\sum_{i=1}^{2}(y_i - \bar{y})^2 = 2(\frac{y_1 - y_2}{2})^2 = (\frac{y_1 - y_2}{\sqrt{2}})^2$ ，也就是說 $(y_1 - \bar{y})^2$ 與 $(y_2 - \bar{y})^2$ 這兩個數值的和，可以完全由 $\frac{y_1 - y_2}{\sqrt{2}}$ 這個數值的平方來決定。因此我們說統計量 $\sum_{i=1}^{2}(Y_i - \bar{Y})^2$ 的自由度為 1。

讓我們再換個角度來作探討。由於

$$\mathbf{y}_2^* = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \end{bmatrix} = \frac{y_1 - y_2}{\sqrt{2}} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

這告訴我們對於任意給定的 y_1, y_2 ， $\mathbf{y}_2^*$ 其實就是具有 $c\ \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ 形式的一個向量，其中 $c \in R$ 。因此儘管 $\mathbf{y}_2^* \in R^2$（亦即 $\mathbf{y}_2^*$ 是二度空間中的一個點），然而這些點並不是任意散佈在平面上，而是全數落於通過 $\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ 的直線上(參見上圖)；也就是說 $\mathbf{y}_2^*$ 其實就是集合

$$\left\{ c\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix} \middle| c \in R \right\},$$

中的一個元素。由於這是一個由單位向量 $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ 所衍生出來的一個子空間，而且這個子空間實際上只是平面上一條通過原點的直線，因此子空間 E_2 的維度等於 1。再者，我們也可以發現由於向量 $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ 滿足 $e_1 + e_2 = 0$ 的這項條件，因此前述集合其實就是 $E_2 = \{\mathbf{e}_2 = (e_1, e_2)' \mid e_1 + e_2 = 0\}$ ，這是一個由平面上所有滿足 $e_1 + e_2 = 0$ 的點所構成的集合。由於 $e_1 + e_2 = 0$ 是個直線方程式，因此 E_2 的維度自然就是 1 了。

同理，就 $n = 3$ 的情形， $\sum_{i=1}^{3}(y_i - \bar{y})^2$ 其實就是 $\mathbf{y}_3^* = (y_1 - \bar{y}, y_2 - \bar{y}, y_3 - \bar{y})'$ 這個向量的長度平方。這個向量有什麼特性呢？首先我們可以發現就任意的 y_1, y_2 和 y_3 來說， $\mathbf{y}_3^*$ 可以分解成

$$\mathbf{y}_3^* = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_3 - \bar{y} \end{bmatrix} = \begin{bmatrix} y_1 - \dfrac{y_1+y_2+y_3}{3} \\ y_2 - \dfrac{y_1+y_2+y_3}{3} \\ y_3 - \dfrac{y_1+y_2+y_3}{3} \end{bmatrix} = \begin{bmatrix} \dfrac{2y_1-y_2-y_3}{3} \\ \dfrac{-y_1+2y_2-y_3}{3} \\ \dfrac{-y_1-y_2+2y_3}{3} \end{bmatrix}$$

$$= \frac{y_1-y_2}{\sqrt{2}}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + \frac{y_1+y_2-2y_3}{\sqrt{6}}\frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \quad 。$$

因此，所有具有$\mathbf{y}_3^*$形式的向量所構成的子空間，其實可以表示成

$$\left\{ \mathbf{y}_3^* \middle| \mathbf{y}_3^* = c_1\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + c_2\frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix},\ c_1, c_2 \in R \right\} 。$$

這是一個由向量$(1,-1,0)'/\sqrt{2}$與$(1,1,-2)'/\sqrt{6}$所衍生出來的一個平面子空間，因此其維度為 2。此外，由於$(1,-1,0)'/\sqrt{2}$與$(1,1,-2)'/\sqrt{6}$是一組互相正交的單位向量，因此向量$\mathbf{y}_3^* = (y_1-\bar{y}, y_2-\bar{y}, y_3-\bar{y})'$的長度平方，實際上可以分解成$\frac{y_1-y_2}{\sqrt{2}}(1,-1,0)'/\sqrt{2}$與$\frac{y_1+y_2-2y_3}{\sqrt{6}}(1,1,-2)'/\sqrt{6}$這兩個正交向量的長度平方和。亦即

$$\sum_{i=1}^{3}(y_i-\bar{y})^2 = c_1^2 + c_2^2$$

$$= (\frac{y_1-y_2}{\sqrt{2}})^2 + (\frac{y_1+y_2-2y_3}{\sqrt{6}})^2$$

因此雖然$\sum_{i=1}^{3}(y_i-\bar{y})^2$是三個項的平方和，但是**本質**上這卻是兩個項(亦即$\sqrt{\frac{1}{2}(y_1-y_2)^2}$與$\sqrt{\frac{1}{6}(y_1+y_2-2y_3)^2}$) 的平方和。因此我們可以說統計量$\sum_{i=1}^{3}(Y_i-\bar{Y})^2$的自由度為 2。

再者，我們也可以發現由於$\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$和$\begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$這兩個向量滿足$e_1+e_2+e_3=0$這個方程式，因此$\mathbf{y}_3^*$自然也需要滿足這項條件。這也就意謂著前述集合其實就是$E_3 = \{\mathbf{e}_3 = (e_1, e_2, e_3)' \mid e_1+e_2+e_3 = 0\}$，這是一個由三度空間裡所有滿足$e_1+e_2+e_3=0$的點所構成的集合。由於$e_1+e_2+e_3=0$是個平面方程式，因此$E_3$的維度自然就是 2 了。

依此類推，就$\mathbf{y}_n^* = (y_1-\bar{y},...,y_n-\bar{y})'$而言，我們可以歸納出下列的結果：

$$\begin{bmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ y_3-\bar{y} \\ y_4-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{bmatrix} = \frac{y_1-y_2}{\sqrt{2}}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \frac{y_1+y_2-2y_3}{\sqrt{6}}\frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots$$

$$+ \frac{y_1+y_2+\cdots+y_{n-1}-(n-1)y_n}{\sqrt{(n-1)n}}\frac{1}{\sqrt{(n-1)n}}\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ -(n-1) \end{bmatrix}$$

其中

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \frac{1}{\sqrt{(n-1)n}}\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ -(n-1) \end{bmatrix}$$

是 $n-1$ 個互相正交的一組向量。因此，所有具有$\mathbf{y}_n^*$形式的向量所構成的子空間，實際上可以表示為

$$\left\{ c_1\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + c_2\frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + c_{n-1}\frac{1}{\sqrt{(n-1)n}}\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ -(n-1) \end{bmatrix} \middle| c_i \in R,\ i=1,2,\ldots,n-1 \right\},$$

其維度為$n-1$。再者由於

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \frac{1}{\sqrt{6}}\begin{bmatrix} 1 \\ 1 \\ -2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \frac{1}{\sqrt{(n-1)n}}\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ -(n-1) \end{bmatrix}$$

這些向量都滿足$\sum_{i=1}^{n} e_i = 0$這個方程式，因此$\mathbf{y}_n^*$自然也需要滿足這項條件。這也就說明了前述集合其實就是$E_n = \left\{ \mathbf{e}_n = (e_1,...,e_n)' \middle| \sum_{i=1}^{n} e_i = 0 \right\}$，這是一個由 n 度空間裡所有滿足$\sum_{i=1}^{n} e_i = 0$的點所構成的集合。由於$\sum_{i=1}^{n} e_i = 0$是個$n-1$度空間方程式，這也再次說明了E_n的維度是$n-1$。此外，既然$\mathbf{y}_n^*$可以分解$n-1$個正交向量的和，因此$\sum_{i=1}^{n}(y_i - \bar{y})^2$自然也可以表示成這$n-1$個正交向量的長度平方和：

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = c_1^2 + c_2^2 + \cdots + c_{n-1}^2$$

$$= \left(\frac{y_1 - y_2}{\sqrt{2}}\right)^2 + \left(\frac{y_1 + y_2 - 2y_3}{\sqrt{6}}\right)^2 + \cdots + \left(\frac{y_1 + y_2 + \cdots + y_{n-1} - (n-1)y_n}{\sqrt{(n-1)n}}\right)^2$$

也就是說$\sum_{i=1}^{n}(y_i - \bar{y})^2$**實質**上是$n-1$個項目的平方和。**因此無論是透過子空間$E_n$的維度數，或者是藉由$\sum_{i=1}^{n}(y_i - \bar{y})^2$實質上所蘊含的項數個數，我們都可以推得統計量$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$的自由度為$n-1$。因此當我們想要了解這$n-1$個項的平均平方值時，自然應該除以$n-1$，而不是$n$了。**