樣本變異數的公式為何是除以 n-1？

淡江大學數學系　鄭惟厚教授

高中數學課本中提到，假設母體數據 $x_1, x_2, \cdots, x_n$ 的平均數等於 μ ，則母體變異數爲 $\sigma^2 = \dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$ ，標準差爲 $\sigma = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$ 。但是如果 $x_1, x_2, \cdots, x_n$ 是抽自某一母體的樣本數據，則其樣本變異數等於 $s^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ ，標準差 $s = \sqrt{s^2} = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ 。同樣都是在計算變異數，爲何一個公式是除以 n，另一個公式卻是除以 n-1 呢？除以 n 很自然，因爲通常求平均都是除以 n 的，所以母體變異數的公式相當「正常」，相對來說，樣本變異數的分母 n-1，就顯得怪怪的。

要了解 n-1 的背後原因，首先必須知道，母體和樣本的角色不同、「地位」也就不一樣。母體是我們的關心對象，通常我們會想要知道有關它的資訊。比如我們可能有興趣知道，全國成年民眾當中，贊成把集會遊行改成報備制的，占多少百分比。但是就如同這個例子的「全國成年民眾」，母體通常都很龐大，幾乎不可能對它蒐集完整資訊，通常只能從其中抽取樣本，再從樣本裡面找相關資訊。

樣本變異數 s^2 的角色，除了可以提供我們有關樣本數據的散佈情況之外，還有一個重要功能，就是當作母體變異數 σ^2 的估計。當作估計的量，我們會希望它不要系統性的高估、或者系統性的低估，也就是要求估計量有「不偏」性質。用秤體重來比喻的話，如果體重計有時把我們秤重了些、有時又秤輕了些，但是若秤了許許多多次之後，平均起來就等於我們的真實體重的話，就相當於有不偏性質。但是假如我家體重計無法正確歸零，常常把我的體重「加碼」，量很多次下來，平均把我多秤了半公斤，這樣就叫做系統性高估，而非「不偏」了。

假設樣本變異數 s^2 公式的分母是用 n 而非 n-1 的話，如果把它當作母體變異數 σ^2 的估計，常常會低估，不符合「不偏」的條件，但是如果把 n 改成 n-1，樣本變異數就會是母體變異數的不偏估計，這件事實是可以用數學證明出來的。當然目前在高中階段，不可能證明給學生看，空口說「不偏」，也有點兒抽象；但是如果用類似以下的例子說明，例子裡面有具體的計算結果，相信學生的接受度會高很多，電腦強的同學，甚至還可以自己驗證。

爲了計算方便，假設母體爲{1, 2, 3,⋯, 1000}，則母體變異數 σ^2 會等於 83333.25。現在假設從母體抽樣，樣本大小是 n = 3，取出不放回；考慮所有可能抽到的樣本，並用兩種不同公式（分別以 n 或 n-1 爲分母）計算樣本變異數，會得到以下結果：

樣本	$\frac{1}{3}\sum_{i=1}^{3}(x_i-\bar{x})^2$	$\frac{1}{2}\sum_{i=1}^{3}(x_i-\bar{x})^2$
1, 2, 3	0.66667	1
1, 2, 4	1.55556	2.33333
1, 2, 5	2.88889	4.33333
.	.	.
.	.	.
.	.	.
998, 999, 1000	0.66667	1
平均	55611.11134	83416.66701

從以上算出的結果可以看出，如果樣本變異數用 n-1= 2 當作分母，所有樣本變異數的平均，即 83416.66701，非常接近母體變異數σ^2 = 83333.25（誤差的主要原因應是取出不放回，稍後會討論），而用 n= 3 當分母的所有樣本變異數，平均等於 55611.11134，就嚴重低估了σ^2 。

樣本變異數用 n-1 當作分母，理應符合不偏性質，爲何所有樣本變異數的平均，和母體變異數σ^2 有一點差距呢？因爲在執行取出不放回時，前後的結果之間並不獨立，而不偏性質是植基於隨機樣本，隨機樣本的各「成員」之間是互相獨立的。爲了印證這一點，可以重新計算一次，而這次的抽樣方式，改爲取出放回。

現在樣本變成(1, 1, 1), (1, 1, 2), (1, 1, 3),…,(1000, 1000, 1000)，重複之前的步驟，替每個樣本計算變異數（分別用 2 或 3 做分母），再求平均可得取出放回之結果如下：

	$\frac{1}{3}\sum_{i=1}^{3}(x_i-\bar{x})^2$	$\frac{1}{2}\sum_{i=1}^{3}(x_i-\bar{x})^2$
平均	55555.50023	83333.25034

用 n= 3 當分母的平均結果，仍然嚴重低估，但是用 n-1= 2 當作分母的平均結果，就等於母體變異數了（小數第四位之後的誤差，應是計算過程中的四捨五入造成的）。

通常在抽樣時，比如做民意調查，實際執行的過程應該相當於取出不放回，因爲同一個人不會被訪問兩次，如此狀況不符合「互相獨立」的條件，是否樣本變異數的不偏性質就不成立了呢？其實只要母體比樣本大很多，取出不放回和取出放回的差別就非常小，即使執行取出不放回，也可以把前後結果視爲互相獨立，舉例說明如下：

假設有 10,000 個人，其中 60%（即 6000 人）會用電腦，現在從母體當中依序抽出三人，則三人都會用電腦的機率，是$\frac{6000}{10000}\cdot\frac{5999}{9999}\cdot\frac{5998}{9998}=0.60000\cdot0.59996\cdot0.59992=0.21596$，這和取出

放回的結果，$0.6 \cdot 0.6 \cdot 0.6 = 0.216$，是非常接近的。但是如果總共只有 10 個人，其中 60%（即 6 人）會用電腦，則依序抽出三人、該三人都會用電腦的機率，會等於

$\frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} = 0.60000 \cdot 0.55556 \cdot 0.50000 = 0.16667$，這就和取出放回的結果，$0.6 \cdot 0.6 \cdot 0.6 = 0.216$，相差很多了。觀察一下就會發現，在這類計算機率的式子當中，最後一項會等於 $\frac{pN - n + 1}{N - n + 1}$，其中的 N 代表母體人數，p 代表母體符合某特質的比例，n 代表樣本大小。假如 N 和 pN 都比 n 大很多，則 $\frac{pN - n + 1}{N - n + 1}$ 會很接近 $\frac{pN}{N} = p$，例如當 $N = 1{,}000{,}000, n = 1{,}000, p = 0.6$ 時，，

$\frac{pN - n + 1}{N - n + 1} = \frac{600000 - 1000 + 1}{1000000 - 1000 + 1} = \frac{599001}{999001} = 0.59960$，而當其他條件不變、$n = 100$ 時，

$\frac{pN - n + 1}{N - n + 1} = \frac{600000 - 100 + 1}{1000000 - 100 + 1} = \frac{599901}{999901} = 0.59996$；既然最後一項都很接近 p，之前的項當然更接近 p 了，所以計算出來的結果就會和取出放回差不多，也就是說，可將依次抽取的各次之間視爲互相獨立。

綜合以上的討論可知，只要母體比樣本大很多的時候，即使樣本是用取出不放回的方式抽取，用 n-1 爲分母的樣本變異數，仍然可視爲母體變異數的不偏估計。