標準差與標準化

卓永鴻 提供

1 標準差

我們經常使用平均數來大致了解一組資料，例如平均成績、平均身高、平均壽命等等。但是如果只看平均數，不見得能足夠了解全體情況。比方說你和郭台銘住同一個社區，你們社區平均每戶年收入兩千萬，那麼你家是有錢還不有錢？為了幫助我們更了解全體情況，我們需要更多統計量來予以描述，讓我們更清楚一組資料的全貌。

其中一個統計量，就是在描述數據的離散情況。換句話說，就是描述這組資料的所有數據，是很集中還是很分散。為了描述離散情況，我們第一個想到的是：每一個數據離平均有多遠？如果對於資料 X 裡的每個數據 x_i，我們都將它減掉平均數 μ_X，然後再全部加起來，得到

$$\sum_{i=1}^{n}(x_i-\mu_X)=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\mu_X$$
$$=n\mu_X-n\mu_X=0$$

只是搞笑，得到 0。因為有些數據比平均多，有些比較少，當然會正負相消。為了防止這種情況，我們改先掛絕對值再相加

$$\sum_{i=1}^{n}\left|x_i-\mu_X\right|$$

這樣就沒有正負相消的問題了。這式子中的每一項，叫做**離均差**，意即每個數據離平均有多遠。然而這樣挺不好算的，一大堆絕對值加在一塊這是很難處理的。所以我們改為先把每個離均差都平方，然後再相加

$$\sum_{i=1}^{n}(x_i-\mu_X)^2$$

可是這個式子，如果數據越多項，豈不是加起來就越大嗎？比方說一班的考試成績，如果隔壁班和本班有一模一樣的成績分佈，兩班合併一起算，算出來就變兩倍，但離散情況並不變。為了消弭數據數 n 的影響，我們再除以 n，變成計算 離均差平方的平均

$$\frac{\sum_{i=1}^{n}(x_i-\mu_X)^2}{n}$$

這個式子，便適合用來描述離散情況，這稱為**變異數** (variance)。但因為我們有先做過平方，所以這樣算出的結果，單位會與原來不一致。如果我們希望單位要一致，便將變異數再開根號

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu_X)^2}{n}}$$

離均差平方的平均再開根號，這稱為**標準差** (standard deviation)。為什麼剛剛說掛絕對值相加比較不好算，平方後再相加卻反而好算呢？我們可以將標準差定義中平方的部份給乘開

$$\begin{aligned}
\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu_X)^2}{n}} &= \sqrt{\frac{\sum_{i=1}^{n}(x_i^2-2\mu_X x_i+\mu_X^2)}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}2\mu_X x_i+\sum_{i=1}^{n}\mu_X^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}x_i^2-2\mu_X\sum_{i=1}^{n}x_i+\mu_X^2\sum_{i=1}^{n}1}{n}} \quad \text{與足碼 } i \text{ 無關的可提出} \\
&= \sqrt{\frac{\sum_{i=1}^{n}x_i^2-2\mu_X(n\mu_X)+\mu_X^2\cdot n}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}x_i^2-2n\mu_X^2+n\mu_X^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}x_i^2-n\mu_X^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n}-\mu_X^2}
\end{aligned}$$

這樣便得到標準差的另一種公式，當我們只知數據平方和而不知道每一組數據的詳細數值，便可以使用此式。

注意

1. 變異數的符號，可寫為 Var 或 σ^2，其公式為

$$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\mu_X)^2}{n}=\frac{\sum_{i=1}^{n}x_i^2}{n}-\mu_X^2$$

2. 標準差的符號，可寫為 $\sqrt{Var}$ 或 σ，其公式為

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu_X)^2}{n}}=\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n}-\mu_X^2}$$

3. 變異數可以想成求算正方形面積的平均，如下圖。數據若離平均越遠，算出的正方形面積就越大，會使變異數算出來更大。

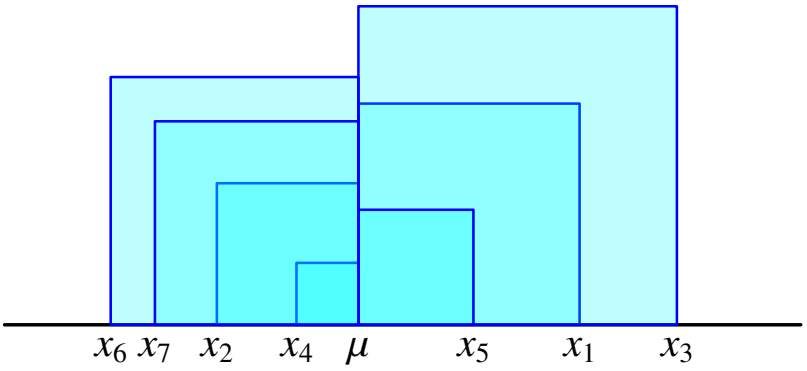

Figure 1: 變異數的幾何意義

4. 資料的平移不影響標準差，但伸縮會影響。設 $Y = aX + b$，則

$$\begin{aligned}
\sigma_Y &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \mu_Y)^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}\Big((ax_i + b) - (a\mu_X + b)\Big)^2}{n}} \\
&= \sqrt{\frac{\sum_{i=1}^{n}\Big(a(x_i - \mu_X)\Big)^2}{n}} \\
&= \sqrt{\frac{a^2\sum_{i=1}^{n}(x_i - \mu_X)^2}{n}} \\
&= |a|\sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_X)^2}{n}} \qquad \sqrt{a^2} = |a| \\
&= |a|\,\sigma_X
\end{aligned}$$

5. 由變異數的第二個式子移項，可得

$$\begin{aligned}
\sigma^2 + \mu_X^2 &= \frac{\sum_{i=1}^{n} x_i^2}{n} \\
\Rightarrow \sum_{i=1}^{n} x_i^2 &= n\left(\sigma^2 + \mu_X^2\right)
\end{aligned}$$

這是個好用的公式。

2 標準化

如果你上次數學考 70 分，這次考 50 分，那麼你是進步還是退步呢？當然，你父母很可能會不太高興，竟然退步了 20 分！正當要執行家法時，你急忙提出抗辯：「等一下！上次數學全班平均 60 分，這次考比較難，班平均只有 40 分！我這兩次都高於平均 10 分，應該不算退步吧！」你父母聽了覺得挺有道理，正當你鬆了口氣時，一旁在讀大學的姐姐看不下去了：「這兩次我都看過你班上成績單，上次標準差才 5 分，你比平均高兩個標準差。這次標準差 10 分，你只比平均高一個標準差而已，所以你還是退步了。」於是你還是被執行家法並回房好好讀數學。當你正生氣覺得不諒解，本來今天高高興興，為什麼姊姊要這麼說這種話，此時剛好翻到數學課本上談到數據的標準化，好像就和姊姊那番話有關係。

課本中寫道：小明身高 187 公分，家庭年收入 1000 萬元。因為小明並不帥，所以顯然他不是高富帥，但究竟說他高比較好呢，還是說他富比較好呢？身高與家庭年收入，顯然是兩種不同的資料，無法直接比較。但是人性嘛，總想比一比，不能比也要比。於是採取相對比較的辦法，小明班上平均身高 175 公分，標準差 4 公分，所以小明身高在班上是比平均多出三個標準差；班上家庭年收入平均 800 萬元，標準差是 200 萬元，所以小明家年收入在班上是比平均多出一個標準差。這樣看來，小明的富在班上看起來好像不是太突出，倒是身高相對來說在班上是比較高。所以我們有了結論：小明(在這個班) 是高而不是富！

為了方便求算比平均高幾個標準差，我們就將數據做個轉換。使轉換後的新數據具有平均數為 1、標準差為 0 的特性，這樣我們只要一看新數據就馬上知道是比平均高多少個標準差了。而要怎麼設定數據轉換呢？當然是這樣寫：

$$Z = \frac{X - \mu_X}{\sigma_X}$$

我們知道對於線性變換 $Y = aX + b$，平均數

$$\mu_Y = a\mu_X + b$$

直接將原平均代入線性變換式子當中。所以

$$\mu_Z = \frac{\mu_X - \mu_X}{\sigma_X} = 0$$

又標準差

$$\sigma_Y = |a|\,\sigma_X$$

與平移無關，伸縮係數掛絕對值。所以

$$\sigma_Z = \left|\frac{1}{\sigma_X}\right| \sigma_X = 1$$

定義 **數據的標準化**

對於資料 X 進行線性變換

$$Z = \frac{X - \mu_X}{\sigma_X}$$

這稱為數據的標準化，數據 Z 稱為標準化數據，標準化數據的值稱為 z 分數。

小明身高 187，進行標準化，就是

$$z = \frac{187 - 175}{4} = 3$$

z 分數是 3，表高出平均 3 個標準差。小明的同學阿花，他是男的，但很愛吃豆花，所以叫阿花。阿花身高 173，進行標準化就是

$$z = \frac{173 - 175}{4} = -0.5$$

z 分數是 -0.5，表比平均低 0.5 個標準差。

標準化數據還有個特性，它是無單位的。比方說小明班上身高，標準差是 4 公分。如果將班上身高單位改用公尺，則每個人數據都變百分之一，例如小明是 1.87 公尺，而班上身高標準差是 0.04 公尺。至於小明身高的 z 分數，無論原數據是用 187 公分還是 1.87 公尺，算出來都是 3。

再介紹一個與標準化概念差不多，但不是直接用 z 分數的例子。我們所談論的 IQ，並不是做完智商測驗後的原始分數。是先將原始分數按不同年齡分類，同年齡的全球平均分數設定為 IQ100，並設定標準差為 15。如果你的原始成績比同年齡全球平均高兩個標準差，那麼你的智商就是 130；如果原始成績比同年齡全球平均低 1.2 個標準差，那麼你的智商就是 82。所以由你做出來的智商，你就可以知道自己在同年齡中的相對高低。假設你 8 歲做一次智商測驗，到 12 歲又做一次，此期間你的智能完全沒有長進也沒有衰退，那麼你的智商應該是會下降的，因為同年齡全體成長了。

有個國際組織叫做 MENSA，是一個高智商俱樂部，其入會門檻為 IQ130。別以為聽起來好像不怎麼高，這已經比平均高兩個標準差，符合的比例僅有約 2.5%[1]！

[1]這比例怎麼來的？高三的機率統計 (二) 會學到如何計算。