Sample variance and population variance

Created by Mr. Francis Hung on 20080906 Last updated: June 17, 2018

有舊生電郵問，為何 sample standard deviation = $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ 而不是 $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$ 呢？

這個問題很複雜，首先，要明白何謂 unbiased estimator。

為了簡單起見，試舉一例如下：有一市鎮，居民三萬人，有一萬人(10^4)每人有一元，另一萬人每人有二元，其餘一萬人每人有六元。由此可求得該批市民的 population mean (μ)，和 population variance (σ^2) 了。

$$\mu = \frac{10^4\times1+10^4\times2+10^4\times6}{3\times10^4} = \frac{1+2+6}{3} = 3 \ldots (1)$$

$$\sigma^2 = \frac{10^4\times(1-3)^2+10^4\times(2-3)^2+10^4\times(6-3)^2}{3\times10^4} = \frac{(-2)^2+(-1)^2+3^2}{3} = \frac{14}{3} \ldots (2)$$

此三萬人為 'Population'。假如我們只知道這市鎮有三萬人，而不知道每人有多少錢，最好的方法，當然是派人員到每家每人去登記；但是所花費的人力物力太大了。我們唯有去取一些樣本(Sample)。比如每個調查員去詢問三人，我們便可以得到許多不同的 Sample，而 Sample Size 為 3。這些樣本的平均值 (sample mean) 為 $\bar{x}$，樣本方差(sample variance)為θ^2。那麼，如何從 $\bar{x}$ 和 θ^2 去估算 μ 和 σ^2 呢?

如果我們派 *n* 個調查員得到 *n* 個樣本的平均值 $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ 和 *n* 個樣本的樣本方差 $\theta_1^2, \theta_2^2, \ldots, \theta_n^2$，則我們可以合理地以 $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ 的平均值和 $\theta_1^2, \theta_2^2, \ldots, \theta_n^2$ 的平均值作為 μ 和 σ^2 的估算。即 $\mu = \frac{\bar{x}_1+\bar{x}_2+\cdots+\bar{x}_n}{n}$，$\sigma^2 = \frac{\theta_1^2+\theta_2^2+\cdots+\theta_n^2}{n}$。但是這是否合理呢？讓我們比較兩者的差異。

如果以 $\theta^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$ 和 $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$。假如我們要取 Sample 有 3 個 elements *a, b, c*，*a, b, c* 可以是 1, 2 或 6，則共有 3×3×3 = 27 個不同的 Sample。

n = 3, $\theta^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2}{3}$, $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2}{2}$

下表可比較不同 *x, y, z* 時，$\bar{x}$, θ^2, s^2 的值：

No. of sample	*a*	*b*	*c*	$\bar{x}$	θ^2	s^2
1	1	1	1	1	0	0
2	1	1	2	$\frac{4}{3}$	$\frac{2}{9}$	$\frac{1}{3}$
3	1	1	6	$\frac{8}{3}$	$\frac{50}{9}$	$\frac{25}{3}$
4	1	2	1	$\frac{4}{3}$	$\frac{2}{9}$	$\frac{1}{3}$
5	1	2	2	$\frac{5}{3}$	$\frac{2}{9}$	$\frac{1}{3}$
6	1	2	6	3	$\frac{14}{3}$	7
7	1	6	1	$\frac{8}{3}$	$\frac{50}{9}$	$\frac{25}{3}$
8	1	6	2	3	$\frac{14}{3}$	7
9	1	6	6	$\frac{13}{3}$	$\frac{50}{9}$	$\frac{25}{3}$

10	2	1	1	$\frac{4}{3}$	$\frac{2}{9}$	$\frac{1}{3}$
11	2	1	2	$\frac{5}{3}$	$\frac{2}{9}$	$\frac{1}{3}$
12	2	1	6	3	$\frac{14}{3}$	7
13	2	2	1	$\frac{5}{3}$	$\frac{2}{9}$	$\frac{1}{3}$
14	2	2	2	2	0	0
15	2	2	6	$\frac{10}{3}$	$\frac{32}{9}$	$\frac{16}{3}$
16	2	6	1	3	$\frac{14}{3}$	7
17	2	6	2	$\frac{10}{3}$	$\frac{32}{9}$	$\frac{16}{3}$
18	2	6	6	$\frac{14}{3}$	$\frac{32}{9}$	$\frac{16}{3}$
19	6	1	1	$\frac{8}{3}$	$\frac{50}{9}$	$\frac{25}{3}$
20	6	1	2	3	$\frac{14}{3}$	7
21	6	1	6	$\frac{13}{3}$	$\frac{50}{9}$	$\frac{25}{3}$
22	6	2	1	3	$\frac{14}{3}$	7
23	6	2	2	$\frac{10}{3}$	$\frac{32}{9}$	$\frac{16}{3}$
24	6	2	6	$\frac{14}{3}$	$\frac{32}{9}$	$\frac{16}{3}$
25	6	6	1	$\frac{13}{3}$	$\frac{50}{9}$	$\frac{25}{3}$
26	6	6	2	$\frac{14}{3}$	$\frac{32}{9}$	$\frac{16}{3}$
27	6	6	6	6	0	0
Total				81	84	126
Average ($\frac{\text{total}}{27}$)				3	$\frac{28}{9}$	$\frac{14}{3}$

從表中最後一行與 (1), (2) 比較，Expected value of $\bar{x}$ is $E(\bar{x}) = \mu = 3$

Expected value of $\bar{x}$ is $E(s^2) = \sigma^2 = \frac{14}{3}$.

而 $\theta^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ 是一個 biased estimator of σ^2 (因為 $E(\theta^2) = \frac{28}{9} \neq \frac{14}{3} = \sigma^2$)。

孔德偉老師