DOI:10.13758/j.cnki.tr.1984.02.011

标准差在地学研究中某些特定条件下的计算方法

刘 多 森

（中国科学院南京土壤研究所）

反映变量离散程度的最常用的统计参数，是方差和标准差。样本的标准差，就是样本方差的平方根。方差和标准差在地学研究中的应用十分广泛。本文将叙述在研究工作中特别是地学研究中的某些特定条件下标准差的计算方法。

众所周知，n 个原始数据 x_i $(i=1,2,\cdots\cdots,n)$ 的标准差 s 为：

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

在不致发生混淆的情况下，也可简记为：

$$s=\sqrt{\frac{\Sigma x^2-\frac{1}{n}(\Sigma x)^2}{n-1}}$$

在研究工作中，经常会遇到文献中只列出样本的容量、均值、标准差而不列出原始数据的情况。这时，如果要把不同文献的相同观测项目（譬如都是土壤铬含量）合并统计，则以各文献的样本容量为权，用计算加权平均值的方法极易求出合并后的均值；但合并后的标准差不能用计算加权平均值的方法求出。

在不知道原始数据的条件下，仅根据分组样本的统计参数——分组的容量、均值、标准差，如何求解合并样本的标准差，可通过如下实例说明。

笔者曾参加中国科学院土壤背景值协作组工作。从这项工作得知：北京地区49个（n_1）土壤标本铬含量（以全Cr计）的均值 $\bar{x}_1$ 为61.52ppm，标准差 s_1 为16.04ppm；南京地区59个（n_2）土壤标本铬含量的均值 $\bar{x}_2$ 为66.74ppm，标准差 s_2 为21.43ppm。显然，将二地区合并统计时，合并样本容量即土壤标本总数 n 为

$$n=n_1+n_2=49+59=108$$

合并样本的均值 $\bar{x}$ 也极易求出：

$$\bar{x}=\frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$$

$$=\frac{49\times61.52+59\times66.74}{108}=64.37\ (\text{ppm})$$

但是，合并样本的标准差 s 不能用类似方法计算，即

$$s\neq\frac{n_1s_1+n_2s_2}{n_1+n_2}$$

正确的计算方法可以这样导出：

$$s^2=\frac{\sum x_1^2+\sum x_2^2-\frac{(\sum x_1+\sum x_2)^2}{n_1+n_2}}{n_1+n_2-1}=\frac{\sum x_1^2+\sum x_2^2-n\bar{x}^2}{n-1}$$

而

$$(n_1-1)s_1^2+(n_2-1)s_2^2=\sum x_1^2-\frac{(\sum x_1)^2}{n_1}+\sum x_2^2-\frac{(\sum x_2)^2}{n_2}$$

$$\sum x_1^2+\sum x_2^2=(n_1-1)s_1^2+(n_2-1)s_2^2+n_1\bar{x}_1^2+n_2\bar{x}_2^2$$

所以

$$s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2+n_1\bar{x}_1^2+n_2\bar{x}_2^2-n\bar{x}^2}{n-1}$$

故北京、南京地区108个土壤标本铬含量的标准差 s 为

$$s=\sqrt{\frac{(49-1)16.04^2+(59-1)21.43^2+49\times61.52^2+59\times66.74^2-108\times64.37^2}{108-1}}$$

$$=19.27\ (\text{ppm})$$

推广到 k 个组，合并样本的标准差 s 为

$$s=\sqrt{\frac{\sum_{i=1}^{k}(n_i-1)s_i^2+\sum_{i=1}^{k}n_i\bar{x}_i^2-n\bar{x}^2}{n-1}}$$

式中 n_i，$\bar{x}_i$，s_i 分别表示第 i 组($i=1,2,\cdots\cdots,k$)的样本容量、均值、标准差；n，$\bar{x}$ 分别表示 k 个组合并后的样本容量和均值。

在地学研究中，有时不必以分组样本的容量为权计算合并样本的统计参数，而以分布面积%数为权进行计算更符合地学思想。这样计算合并均值时，可用分布面积百分数为权，按计算加权平均值的方法极易解出。但计算合并标准差时，如果直接用各组面积%数 a_i 作为相应组的样本容量 n_i 进行计算，则因 a_i 与1之间的差异不够大，以致影响了计算结果的精度。因此，在实际工作中，为了减少由 (a_i-1) 引起的误差，可将各组面积百分数 a_i 都乘以100(或1000)，用这个乘积作为相应组的样本容量 n_i，用100×100(或100×1000)作为合并样本容量 n 进行计算。这样，合并样本的标准差 s 一般为

表 1　全太湖地区水稻土有机质的均值和标准差以面积为权的计算结果

水稻土类型	面积 a_i (%)	有机质含量(%)	
		均值 $\bar{x}_i$	标准差 s_i
爽水水稻土	29.2	2.81	0.545
渗水水稻土	13.3	1.59	0.418
滞水水稻土	8.3	2.23	0.561
囊水水稻土	25.9	3.40	1.28
漏水水稻土	23.3	2.02	0.599
全地区	$a=100.0$	$\bar{x}=2.57$	$s=1.02$

注：摘引自参考文献[1]。

$$s=\sqrt{\frac{\sum_{i=1}^{k}(100a_i-1)s_i^2+100\sum_{i=1}^{k}a_i\bar{x}_i^2-10000\bar{x}^2}{10000-1}}$$

以表1为例，合并均值即全地区水稻土有机质含量的均值 $\bar{x}$，可用各类水稻土面积百分数为权进行计算：

$$\bar{x}=\frac{1}{100}(29.2\times2.81+13.3\times1.59+8.3\times2.23+25.9\times3.40+23.3\times2.02)=2.57\ (\%)$$

而当计算合并标准差即全地区水稻土有机质含量的标准差 s 时，可将分组样本容量 n_i 视为该组水稻土面积百分数 a_i 乘以100，而将合并样本容量 n 视为100×100=10000，则合并标准差 s 为

$$s=\sqrt{\frac{\sum_{i=1}^{5}(100a_i-1)s_i^2+100\sum_{i=1}^{5}a_i\bar{x}_i^2-10000\times2.57^2}{10000-1}}=1.02(\%)$$

文献[1]有关表中全太湖地区水稻土的有机质、全氮、全磷、速效磷、全钾、速效钾含量的标准差，正是按这种方法从五个类型水稻土分布面积%数及有关统计参数计算的。

参　考　文　献

[1] 徐琪、陆彦格、刘元昌、朱洪官：中国太湖地区水稻土。67～69页，上海科学技术出版社，1980。