

土壤普查资料计算平均值标准差和方差分析的方法

陈晓南

(浙江丽水市农业局)

摘 要

简介了在整理土壤普查资料中,如何以大量的原始数据求取 $\bar{X}$ (总), S (总)和方差分析的简便方法。

本方法主要是利用已知的 n , $\bar{X}$, S 合并计算 $\bar{X}_{总}$ 、 $S_{总}$ 和方差分析,而不是直接从大量的成百上千的原始数据直接进行计算,因而达到了减少计算工作量、避免计算和操作误差并提高汇总数据准确度的目的。

下面分别介绍几个方法:

一、 $\bar{X}_{总}$ 、 $S_{总}$ 的计算方法^{〔1〕}

$\bar{X}_{总} = \frac{\sum f_i \bar{X}_i}{\sum f_i}$ (加权平均数法,式中 f_i 为各样本 n 值, $\bar{X}_i$ 为各样本平均数)

$$S_{总} = \sqrt{\frac{\sum_{i=1}^k (n_i - 1) S_i^2 + \sum_{i=1}^k n_i \bar{X}_i^2 - n_{总} \bar{X}_{总}^2}{n_{总} - 1}}$$

推导从略。以下举例。

示例:松阳、遂昌和龙泉三个县的黄泥沙田土有机质含量的原始数据和以县为单位的 n , $\bar{X}$ 、 S 值列于表1。土壤共有258个样品,此处仅举76个样品的资料为例。

表1 松阳、遂昌和龙泉三县黄泥沙田有机质含量(%)

松 阳	2.24	2.60	2.34	2.69	2.41	3.01	1.88	2.60	3.53			
	n = 12		X = 2.59		S = 0.409							
遂 昌	2.33	2.89	3.57	3.35	2.92	2.22	3.08	3.73	3.74	2.68	3.73	
	3.00	2.98	3.13	n = 14	X = 3.10		S = 0.49					
龙 泉	4.06	2.74	3.86	4.15	4.07	3.32	5.92	3.86	6.33	5.16	4.87	4.31
	3.71	2.39	5.10	3.21	3.54	3.55	3.92	3.12	3.95	4.13	4.71	4.60
	3.58	3.19	3.46	3.28	3.36	3.48	3.96	4.13	4.45	3.42	3.88	3.27
	3.52	4.66	3.78	3.85	3.28	n = 41	X = 3.94		S = 0.781			

利用已知 n 、 $\bar{X}$ 、 S 计算:

$$\bar{X}_{\text{总}} = \frac{2.59 \times 12 + 3.10 \times 14 + 3.94 \times 41}{12 + 14 + 41} = 3.52$$

$$S_{\text{总}} = \sqrt{\frac{(12-1) \times 0.409^2 + (14-1) \times 0.49^2 + (41-1) \times 0.781^2 + 2.59^2 \times 12 + 3.1^2 \times 14 + 3.94^2 \times 41 - 3.52^2 \times 67}{67-1}} = 0.877$$

用经典公式验证计算:

$$\bar{X}_{\text{总}} = \frac{2.24 + \dots + 2.33 + \dots + 3.85}{67} = 3.52$$

$$S_{\text{总}} = \sqrt{\frac{(2.24^2 + \dots + 2.33^2 + \dots + 3.85^2) - \frac{(2.24 + \dots + 2.33 + \dots + 3.85)^2}{67}}{67-1}} = 0.868$$

如果以经典法为正确, 则本方法计算得到的 $S_{\text{总}}$ 误差为: $\frac{|0.887 - 0.868|}{0.868} \times 100 = 1.04\%$ 。这是由小数取位引起的, 是一种允许误差:

计算 $S_{\text{总}}$ 时常会发生两种错误的算法:

(1) $S_{\text{总}} = \frac{\sum S_i n_i}{\sum n_i}$ 上例按此式得 $S_{\text{总}} = 0.654$, 这是错误的结果。

(2) 计算器法计算加权平均数得 $\bar{S}_{\text{总}} = 3.52$ 时, 同时可得 $S = 0.552$, 这也是错误的结果。

二、利用已知 n 、 $\bar{X}$ 、 S 作方差分析的方法

关于利用已知 n 、 $\bar{X}$ 、 S 作方差分析的方法, 笔者在实际工作中摸索总结了两种方法, 为叙述方便, 分别称之为 A 法, B 法。

1. A 法: 利用已知 n 、 $\bar{X}$ 、 S 求得 $\sum X$, $\sum X^2$ 而后作各项平方和及均方的常规计算, 将表 1 中的有关数据列于表 2, 并证明如下:

$$\sum X = n\bar{X}$$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1} \quad \text{移项得:}$$

$$\sum X^2 = S^2(n-1) + \frac{(\sum X)^2}{n}$$

表 2 三个县黄泥沙田有机质含量的 $\sum X$ 、 $\sum X^2$ 求算表

县 别	n	$\bar{X}$	S	$\sum X$	$\sum X^2$
松阳	12	2.59	0.409	31.08	82.34
遂昌	14	3.10	0.49	43.4	137.66
龙泉	41	3.94	0.781	161.54	660.87
合计	67	—	—	236.02	880.87

所以表 2 $\sum X$ 、 $\sum X^2$ 的计算是成立的, 于是有: $C = 236.02^2/67 = 831.42$;

$SS_T = 880.87 - 831.42 = 49.45$, $d_{fT} = 67 - 1 = 66$;

$SS_A = (\frac{31.08^2}{12} + \frac{43.4^2}{14} + \frac{161.54^2}{41}) - 831.4 = 20.08$, $d_{fA} = 3 - 1 = 2$;

$$SS_e = SS_T - SS_A = 29.37, \quad d_{fe} = 66 - 2 = 64$$

由此可得表3。表3表明各县有机质含量差异达极显著水准，可进一步作多重比较，此处从略。

2. B法：A法计算 $\sum X^2$ 时仍嫌较烦，笔者又摸索了更为简捷的B法。其计算公式为：

$$SS_A = S^2 \cdot (n_{\text{总}} - 1); \text{大样本时 } SS_A = S^2 \cdot n_{\text{总}}.$$

式中S为加权平均数的S值。

$$SS_e = \sum S_i^2 (n_i - 1)$$

试对B法的计算公式推导如下：

$$SS_A = \sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1} \Rightarrow S^2 (n - 1) = \sum X^2 - \frac{(\sum X)^2}{n} \Rightarrow SS_A = S^2 \cdot (n - 1)$$

$$\text{大样本时 } S^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}, \text{ 所以 } SS_A = S^2 \cdot n_{\text{总}}$$

有了 SS_A 、 SS_e 即可迅速算出F值，免去了C和 SS_T 的计算，(如需要 SS_T ，则 $SS_A + SS_e$ 即可)。

示例：利用表1中已知 n 、 $\bar{X}$ 、S计算：

$$SS_A = 0.552^2 \times (67 - 1) = 20.08$$

$SS_e = 0.409^2 \times (12 - 1) + 0.49^2 \times (14 - 1) + 0.781^2 \times (14 - 1) = 29.36$ ，进而可得 $F = 21.89$ ，结果与A法基本一致。

3. 用经典法对A、B两法的验证：

根据表1原始数据计算得：

$$\sum X = 235.99, \quad \sum X^2 = 880.88, \quad C = 831.21$$

$$SS_T = \sum X^2 - C = 49.67$$

$$SS_A = \left(\frac{31.03^2}{12} + \frac{43.35^2}{14} + \frac{161.61^2}{41} \right) - 831.21 = 20.28, \quad MS_A = 10.14$$

$$SS_e = 49.67 - 20.28 = 29.39 \quad MS_e = 0.459, \quad F = 22.09$$

如果以经典法为绝对正确(其实也有误差，为验证A、B法，权且将经典法看成绝对正确。)则：A法误差 $\frac{|21.87 - 22.09|}{22.09} \times 100 = 0.996\%$ ，B法误差 $\frac{|21.89 - 22.09|}{22.09} \times 100$

$= 0.905\%$ ，可见A、B法的误差是很小的，属于允许误差范围。

4. A、B法的实用价值

(1) 解决了许多资料只列出 n 、 $\bar{X}$ 、S值而不列原始数据，以至无法作方差分析的问题，从而可使研究工作深入进行。(2) 可将大样本分解成若干小样本，分别求得 n 、 $\bar{X}$ 、S后再用

表 3 方差分析表

变异来源	自由度	平方和	均 方	F
样本间	2	20.08	10.04	21.87**
误 差	64	29.37	0.459	
总变异	66	49.45		

$$F_{0.01}(2, 64) = 4.95$$

B法(或A法)作方差分析,这样可减少大样本在输入计算机(或计算器)时的操作错误,从而提高数据的准确性。表4列出黄泥沙田9个县(市)258个样品的各县 n 、 $\bar{X}$ 与 S ,应用B法很快计算得到 $SS_A = 44.34$ (此处用大样本公式), $SS_e = 107.42$,进而算得 $F = 14.7419$ 。

如果用经典法,258个样品的输入过程极易出错,且不易检查。这里为验证B法,笔者计算如下:

$$\Sigma X = 778.11, \Sigma X^2 = 2498.55, n = 258$$

$$C = 778.11^2 / 258 = 2346.73, SS_T = 2498.55 - 2346.73 = 151.82;$$

$$SS_A = \sum \frac{(\Sigma X_i)^2}{n_i} - C = 44.35 \quad SS_e = 151.82 - 44.35 = 107.47, \text{进而算得 } F = 14.7350,$$

所得结果与B法几乎完全一致。

表 4 9县(市)黄泥沙田有机质含量的 $n, \bar{X}, S$ 值

项 目	缙 云	丽 水	青 田	云和、景宁	松 阳	遂 昌	龙 泉	庆 元
n	76	35	40	15	12	14	41	25
$\bar{X}$	2.83	2.77	2.94	2.92	2.59	3.10	3.94	2.75
S	0.523	0.511	0.855	0.739	0.409	0.490	0.781	0.722

表4仅为一个土种一个项目的资料。如果统计对象为土属(例如黄泥沙田土属有机质样品数 $n = 451$)、亚类、土类乃至全土壤的所有项目,成千上万个原始数据,用经典法几乎是无法计算 $\bar{X}_{\text{总}}$ 、 $S_{\text{总}}$ 和对各样本进行方差分析的。

参 考 文 献

- [1]刘多森,标准差在地学研究中某些特定条件下的计算方法,土壤,16卷第2期,1984。