

方差分析法浅析——单因素的方差分析

杨小勇

(广东石油化工学院 化学与生命科学学院, 广东 茂名 525000)

摘要: 浅析了方差分析法的基本原理和方差分析的意义: 方差分析就是将总的方差分解为各个方差的成分, 然后利用显著性检验法进行分析判断和做出适当的结论; 详细地讲解、介绍了单因素的方差分析法, 回答了学生提出的如何判断一个实验者使用几种不同的方法去测定某种试样的成分或每一参数, 每一种方法都重复测定几次; 如何判断几种不同的方法所得的几个平均值之间是否存在差异, 哪一个平均值准确可靠的问题。

关键词: 方差; 方差分析法; 单因素; 判断

中图分类号: O212.1 **文献标志码:** A **doi:** 10.3969/j.issn.1672-4550.2013.01.015

Discussion of Variance Analysis: the Single Factor Variance Analysis

YANG Xiao-yong

(Department of Chemistry and Life Science, Guangdong University of Petrochemical Technology, Maoming 525000, China)

Abstract: In this paper, the principle and importance of variance analysis are discussed. The variance analysis includes decomposition of general variance into each component part and analyzing and making conclusions by the use of test of significance approach. The single factor variance analysis is explained in detail. Several questions raised by students are also answered including how to judge whether there is a difference among the average values using different methods and which average value is more reliable when a researcher determines the components or parameters of a specimen using different methods.

Key words: variance; variance analysis; single factor; judgement

在实验教学过程中, 常有学生问到这样的问题: 一个实验者使用几种不同的方法去测定某种试样的成分或每一参数, 每一种方法都重复测定几次, 如何判断几种不同的方法所得的几个平均值之间是否存在差异, 哪一个平均值准确可靠。若要回答学生这些问题需运用方差分析法的有关知识。

在科学研究中, 为了探索某一项分析任务的可靠性和影响因素, 需要进行大量的实验^[1]。例如, 取几批试样分别送到几个相关的实验室用不同的方法进行实验, 每一方法的测定又重复若干次, 这样就得到大量数据。根据得到的这些数据, 用什么方法可以判断哪一个因素对测定结果影响最大? 哪一个因素影响不大? 常用的方差分析法就是一种处理和判断数据的手段^[2]。下面结合实例简单浅析方差分析法有关知识。

1 标准偏差和方差^[3]

在数理统计中常用标准偏差表示精密度, 标准偏差(简称标准差)或称均方误差, 用 σ 代表, 表达式为

收稿日期: 2012-04-19; 修改日期: 2012-06-05

作者简介: 杨小勇(1959-), 男, 本科, 高级实验师, 研究方向: 物理化学实验教学。

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad (1)$$

上式是应用于大量数据的条件下(一般有 30 次以上测定), 这时, 测定的平均值接近于真值, 用 μ 表示, 根式的分子代表各别测定数据的偏差的平方和, N 是测定次数。方差代表各偏差平方和的平均值, 用 σ^2 表示。我们知道, 只有当测定次数为无限多或者至少有 30 次以上时, 所得的平均值称真值。但是, 通常测定次数总是有限的, 这样, 所得的平均值并不是真值, 这时, 如果按公式(1)去计算标准偏差就不合理。

测定平均值与真值不相等, 测定值与平均值之差同测定值与真值之差不相等。为要表示有限测定次数的精密度采用符号 S 代替 σ , 有限测定次数标准偏差的计算公式为:

$$S = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} \quad (2)$$

与公式(1)比较, 用试样平均值 $\bar{x}$ 代替了真值 μ , 用 $N-1$ 代替 N 。 $N-1$ 在数理统计中称之为“自由度”, 它说明在 N 次测定中, 只有 $N-1$ 个可变的偏差。自由度也可理解为: 数据中可供对比的数目。例如, 两次测定 a 和 b 只有一个 a 与 b 之间的比较,

三次测定可有两种比较(即其中任何两个数据之间及这两个数据的平均值与第三个数据之间比较),类推下来, N 次测定只有 $N-1$ 个可供对比。如果有几个互相独立的因素同时影响测定结果,设不同来源的方差分别为 $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$,按方差的加和法则,总的方差 σ^2 应为各个方差之和:

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2$$

方差分析就是将总的方差分解为各个方差的成分,然后利用显著性检验法进行分析判断和做出适当的结论^[4]。方差分析有:(1)单因素的方差分析;(2)双因素的方差分析;(3)具有重复试验的双因素方差分析;(4)三因素的方差分析。本文仅讨论单因素的方差分析,即,一个实验者使用几种不同的方法去测定某种试样的成分,每一种方法都重复测定几次,如何判断不同的方法所得的平均值之间是否存在差异。

2 单因素的方差分析方法^[5]

单因素的方差分析是最简单的一种分析,它只有1个能确定的因素可能影响结果。例如,一个实验者使用 k 种不同的方法(在统计分析中也称为 k 个水平)去测定某种试样的成分,每一种方法都重复测定 n 次,如何判断这个 k 种方法的比较试验中所得 k 个平均值之间是否存在差异,这里只有因素就是“方法”。

每一种方法都在相同条件下重复 n 次,测定总是有偶然误差的,假设通过多次测定所得的方差为 σ_0^2 ,对于 k 种方法中的任一方法来说,其自由度应为 $n-1$,对于全部实验的测定数据来说,其自由度应为 $k(n-1) = kn - k$ 。

我们研究的对象是方法这个因素的影响。假如每一种方法都测定无限次,所得的平均值就称真值,那么,从 k 个平均值之间的精密度就可以估量出方法本身固有的方差(以 σ_n^2 表示)。由于每种方法都进行测定 n 次, n 次测定平均值的方差应是 σ_0^2/n 。在计算 k 种方法之间的方差时,应考虑方差的两个来源,一是来自方法本身,另外一个来自实验的偶然误差,按方差的加和法则,得

$$\sigma_1^2 = \sigma_0^2/n + \sigma_n^2$$

式中 σ_1^2 代表方法之间的方差,它的自由度为 $k-1$,假设方法之间没有差别,即每一种方法的总体平均值都相同, $\sigma_n^2 = 0$ 。我们能计算:

$$(1) \sigma_0^2, f_0 = k(n-1);$$

$$(2) \sigma_1^2 = \sigma_0^2/n + \sigma_n^2, f_1 = k-1.$$

将(2)的计算式两边乘以 n ,可得:

$$n\sigma_1^2 = \sigma_0^2 + n\sigma_n^2$$

在显著性检验中,总是假设方法间没有实际差异,即假设 $\sigma_n^2 = 0$,由上式可以认为 $n\sigma_1^2$ 就是 σ_0^2 的估量值。

为了判断方法间是否存在差异,可用 F 检验:

$$F = \frac{\sigma_0^2 + n\sigma_n^2}{\sigma_0^2} = \frac{n\sigma_1^2}{\sigma_0^2} (f_1, f_0) \quad (4)$$

用式(4)计算出来的 F 值与附表(F 检验的临界值)上的值进行比较,有4种情况^[5]:

(1) $F > F_{99\%}$, 方法影响特别显著,记为“*”;

(2) $F_{99\%} \geq F > F_{99.5\%}$, 方法影响显著,记为“*”;

(3) $F_{99.5\%} \geq F > F_{99\%}$, 方法有一定影响,记为“*”;

(4) $F_{90\%} \geq F$, 看不出方法对结果有较大影响。

3 单因素的方差分析方法例题分析^[6]

例题: 一个实验分析者利用静力学原理用4种不同方法在 20.0°C ,常压下测定 $0.005 \text{ c/mol} \cdot \text{L}^{-1}$ 正丁醇的液体表面张力,每种方法的测定均重复5次,所得数据如表1所示,试判断方法之间是否有实际差异。

表1 测定数据 $\sigma / 10^{-3} \text{ N} \cdot \text{m}^{-1}$

方法 A	方法 B	方法 C	方法 D
71.31	71.31	71.32	71.34
71.33	71.31	71.34	71.35
71.34	71.31	71.33	71.33
71.32	71.32	71.35	71.32
71.34	71.32	71.36	71.34

解: 为了简化计算过程,将原数据减去71.30,表1数据转化为表2的数据。

表2 测定数据 $\sigma / 10^{-3} \text{ N} \cdot \text{m}^{-1}$

方法 A	方法 B	方法 C	方法 D
1	1	2	4
3	1	4	5
4	1	3	3
2	2	5	2
4	2	6	4

设方法内部测定数据的离散性用实验误差的方法 σ_0^2 表示。方法之间测定数据的离散性用各种方

法所得平均值之间的方差 σ_1^2 来表示。显著性检验用 F 检验法:

$$F = \frac{\sigma_0^2 + n\sigma_n^2}{\sigma_0^2} = \frac{n\sigma_1^2}{\sigma_0^2}$$

计算 σ_0^2 和 σ_1^2 可用如下 2 式:

$$\sigma_0^2 = \frac{1}{k(n-1)} \sum_1^k \left[\sum_1^n (x - \bar{x})^2 \right] \quad (5)$$

$$\sigma_1^2 = \frac{1}{k-1} \left[\sum \bar{x}_i^2 - \frac{1}{k} \left(\sum \bar{x}_i \right)^2 \right] \quad (6)$$

为了计算 σ_0^2 , 先计算一些数据: $\sum x$, $\sum x^2$, $\frac{1}{5}(\sum x)^2$, $\sum_1^n (x - \bar{x})^2 = \sum x^2 - \frac{1}{5}(\sum x)^2$, 如表 3 所示。

表 3 $\sum x$, $\sum x^2$, ... 数据表

	x_A	x_A^2	x_B	x_B^2	x_C	x_C^2	x_D	x_D^2
	1	1	1	1	2	4	4	16
	3	9	1	1	4	16	5	25
	4	16	1	1	3	9	3	9
	2	4	2	4	5	25	2	4
	4	16	2	4	6	36	4	16
$\sum x$	14		7		20		18	
$\sum x^2$		46		11		90		70
$\frac{1}{5}(\sum x)^2$		39.2		9.8		80.0		64.8
$\sum_1^n (x - \bar{x})^2 = \sum x^2 - \frac{1}{5}(\sum x)^2$		6.8		1.2		10.0		5.2

$$\sigma_0^2 = \frac{6.8 + 1.2 + 10.0 + 5.2}{4(5-1)} = 1.45$$

$$\bar{x}_A = 14/5 = 2.8 \quad \bar{x}_B = 7/5 = 1.4$$

$$\bar{x}_C = 20/5 = 4.0 \quad \bar{x}_D = 18/5 = 3.6$$

f_0 对于全部实验的测定数据来说, 其自由度应为 $k(n-1) = 4(5-1) = 16$

为了计算 σ_1^2 , 先从 $\bar{x}_i$ 求 $\bar{x}_i^2$, 如表 4 所示。

表 4 $\bar{x}_i, \bar{x}_i^2$ 各值

$\bar{x}_i$	2.8	1.4	4.0	3.6	$\sum \bar{x}_i = 11.8$
$\bar{x}_i^2$	7.84	1.96	16.00	12.96	$\sum \bar{x}_i^2 = 38.76$

$$\text{故得: } \sigma_1^2 = \frac{1}{3} \left[38.76 - \frac{(11.8)^2}{4} \right] = 1.317$$

f_1 对于 k 种方法中的任一方法来说, 其自由度应为 $n-1 = 4-1 = 3$ 。 F 检验:

$$F = \frac{n\sigma_1^2}{\sigma_0^2} = \frac{5 \times 1.313}{1.45} = 4.54 \quad (f_0 = 16, f_1 = 3)$$

由 F 附表^[7] (F 检验的临界值) 得

$$F_{99.5\%}(3, 16) = 3.24, F_{99\%}(3, 16) = 5.29$$

由第 2 节 F 检验的临界值 $F_{99\%} \geq F > F_{99.5\%}$, 判断方法之间影响显著, 记为“*”说明方法之间存在差异^[8], 无法判断哪一个平均值准确可靠。至此单因素的方差分析完毕。

参 考 文 献

- [1] 任 弘, 章潮晖, 王利群. 北京体育大学学报 [J]. 2004, 27(3): 359-3603.
- [2] 东北师范大学. 物理化学实验 [M]. 2 版. 北京: 高等教育出版社, 1989: 20.
- [3] 宋清. 定量分析中的误差和数据评价 [M]. 北京: 人民教育出版社, 1982: 14-15.
- [4] 郭太安. 数理统计与管理 [J]. 数理统计与管理, 1989 (6): 36.
- [5] 东北师范大学等. 物理化学实验 [M]. 2 版. 北京: 高等教育出版社, 1989: 21.
- [6] 东北师范大学. 物理化学实验 [M]. 2 版. 北京: 高等教育出版社, 1989: 21-24.
- [7] 宋清. 定量分析中的误差和数据评价 [M]. 北京: 人民教育出版社, 1982: 162-165.
- [8] 宋清. 定量分析中的误差和数据评价 [M]. 北京: 人民教育出版社, 1982: 103.