方差分析

定义

方差分析的目的是用数量的形式来分析导致试验结果不同的两种原因：

- 观测因素：由观测条件不同引起，是系统性差异
- 随机因素：由随机因素干扰引起，是偶然性因素
  确定出在试验中有无由观测条件引起的系统性因素在起作用；从实验数据中分析出各种因素的影响以及其相互间的交互关系，并确定出这些影响作用的大小或程度。

基本概念

- 试验指标：待考察的现象某指标的数量指标（生产玉米的甜度）

- 因素：影响指标的条件
  - 可控因素：所用土地
  - 不可控因素：气象条件，测量误差

- 因素水平：因素所处的状态（用土地 A 种，或者用土地 B 种）

- 单因素试验：试验中只有一个因素在起作用（只有土地不同）
- 多因素试验：试验中有多个因素在起作用（土地和浇水量都不同）

单因素方差分析

设因素水平有 m 个，分别为 $A_1, A_2, \cdots, A_m$ ，在每个因素下分别做了 k_i 次试验，总试验次数为 n ， $n = \sum_{i=1}^{m} k_i$ 。每次实验结果记为 X_{ij} ，表示在第 i 个水平下的第 j 次试验。试验观测值为 x_{ij} 。

因为试验仅有一个因素，且不知道该因素是否真正有影响，故可将 $A_1, A_2, \cdots, A_m$ 看作 m 个总体，于是 $X_{i1}, X_{i2}, \cdots, X_{ik_i}$ 是取自总体 A_i 的一个容量为 k 的样本。

为方便研究进一步假定这些总体为 **正态总体** 且 **具有相同的方差** ，于是：

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, 2, \cdots, m, \quad j = 1, 2, \cdots, k_i$$

原问题：考察因子 A 对实验结果是否有显著的影响，转化为：同方差多（正态）总体期望是否相同。

设 ϵ_i 是因子 A 的第 i 个水平 A_i 所引起的差异：

$$\epsilon_i = \mu_i - \mu$$

称为水平 A_i 的效应，其中 $\mu = \frac{1}{m} \sum_{i=1}^{m} \mu_i$ 称为均值（期望）的总平均。从而问题又可以描述为：推断因子 A 各水平效应之间是否有显著差异。

对于这类问题，产生了两种统计分析问题：

1. 假设检验问题：各个总体的均值是否相等
2. 参数估计问题：对 $\mu_1, \mu_2, \cdots, \mu_m, \sigma^2$ 进行参数估计。

现在关注假设检验问题。原假设：

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

引入新统计量：
组内平均值（水平 A_i 下的样本均值）：

$$\bar{X}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} X_{ij}$$

数据总平均

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k_i}X_{ij}$$

总离差平方和，反映全部数据间差异程度的数量指标：

$$S_T=\sum_{i=1}^{m}\sum_{j=1}^{k_i}(X_{ij}-\bar{X})^2$$

误差平方和，反映组内（统一水平下）样本的随机波动：

$$S_e=\sum_{i=1}^{m}\sum_{j=1}^{k_i}(X_{ij}-\bar{X}_i)^2$$

效应平方和，在一定程度上反映由组间（因子各个不同水平之间）不同而引起的差异：

$$S_A=\sum_{i=1}^{m}\sum_{j=1}^{k_i}(\bar{X}_i-\bar{X})^2$$

同时，易证平方和分解公式：

$$S_T=S_e+S_A$$

该公式说明总体数据的差异可以看作是由两个部分组成：

- 组内平方和（误差平方和），即由随机因素引起的差异。
- 组间平方和，即由因子水平不同引起的差异。

直观上，如果假设成立，各个水平的效应应该是 0，即 S_A 也只含有随机误差，所以 S_A 和 S_e 不应该有太大的差异。试图通过两者的比值来反映随机因素和因子水平不同引起的差异大小，于是构造检验统计量：

$$F=\frac{S_A/(m-1)}{S_e/(n-m)}$$

引理：平方和分解定理

- 设 Q 服从自由度为 n 的 χ^2 分布，又

$$Q_1+Q_2+\cdots+Q_k=Q$$

其中 Q_k 是秩为 f_i 的非负二次型，则 Q_i 相互独立且分别服从自由度为 f_i 的 χ^2 分布的充要条件是

$$f_1+f_2+\cdots+f_k=n$$

已知 H_0 为真时:

$$\frac{S_T}{\sigma^2}=\frac{nS}{\sigma^2}\sim\chi^2(n-1)$$

秩为 $n-1$，其中 S 为样本总方差。S_A/σ^2 和 S_e/σ^2 均为非负二次型，秩分别为 $f_A=m-1$ 和 $f_e=n-1$ 。

- 证明：$S_A=\sum_{i=1}^{m}k_i(\bar{X}_i-\bar{X})^2$ 而 $\bar{X}_m-\bar{X}=-\sum_{i=1}^{m-1}(\bar{X}_i-\bar{X})$，于是 S_A 只有 $m-1$ 个二次项。

由平方和分解定理，我们知道

$$\frac{S_A}{\sigma^2}\sim\chi^2(m-1),\quad\frac{S_e}{\sigma^2}\sim\chi^2(n-m)$$

于是检验统计量满足：

$$F\sim F(m-1,n-m)$$

可证得：

$$E(\frac{S_A}{m-1})=\sigma^2+\frac{k}{m-1}\sum_{i=1}^{m}\epsilon_i^2$$

$$E(\frac{S_e}{m(k-1)})=\sigma^2$$

于是拒绝域为 $F_A>F_{1-\alpha}(m-1,n-m)$

方差分析显著性检验具体步骤

1. 为了方便计算，通常作数据线性变换预处理：

$$y_{ij} = \frac{x_{ij} - c}{d}$$

2. 将新数据列表

温度i / 试验号j	1 40°C	2 50°C	3 60°C	4 70°C	5 80°C	6 90°C	合计
1	4.3	6.1	10.0	6.5	9.3	9.5	
2	7.8	7.3	4.8	8.3	8.7	8.8	
3	3.2	4.2	5.4	8.6	7.2	11.4	
4	6.5	4.1	9.6	8.2	10.1	7.8	
T_i	21.8	21.7	29.8	31.6	35.3	37.5	177.7
n_i	4	4	4	4	4	4	24
$\bar{x}_i$	5.450	5.425	7.450	7.900	8.825	9.375	$\bar{x}$=7.404

3. 按照公式计算
4. 列出方差分析表
5. 根据显著性水平，进行统计检验得出结论。

$$S_T = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}^2 - CT$$

$$S_A = \sum_{i=1}^{m}\left(T_i^2 / n_i\right) - CT$$

$$S_e = S_T - S_A$$

$CT=177.7^2/24=1315.27$

$S_T=(4.3^2+7.8^2+\ldots+7.8^2)-1315.27=112.27$

$S_A=(21.8^2+21.7^2+\ldots+37.5^2)/4-1315.27=56$

$S_e=112.27-56=56.27$

$F=(18\times56)/(5\times56.27)=3.583$

$F_{0.95}(5,18)=2.77,\ F_{0.99}(5,18)=4.25$

显著性水平α下的否定域(拒绝域)

$F_A > F_{1-\alpha}(m-1, m(k-1))$

方差来源	平方和	自由度	F 值	分位数	显著性
温度	56	5	3.583	$F_{0.95}$=2.77 $F_{0.99}$=4.25	*
误差	56.27	18			
合计	112.27	23			

双因子方差分析

双因子间无相互作用的情形

假定 A, B 是影响实验结果的两个因子，因子 A 有 n 个水平：$A_1, A_2, \cdots, A_n$；因子 B 有 m 个水平：$B_1, B_2, \cdots, B_m$。在 A 的 n 个水平与 B 的 m 个水平的每种组合下作一次试验，实验结果以 X_{ij} 表示，其观测值记为 x_{ij}。

假定在每种组合下试验结果 X_{ij} 满足：

$$X_{ij} \sim N(\mu_{ij}, \sigma^2)$$

并设满足如下定义的效应可加性

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

其中 α_i 表示因子水平 A_i 的效应，表示因子 A 的第 i 个水平对试验结果带来的影响，β_j 为因子 B 的效应，表示因子 B 的第 j 个水平对试验结果带来的影响。

问题转化为检验如下假设：

$$H_0:\ \mu_{11} = \mu_{12} = \cdots = \mu_{ij} = \cdots = \mu_{nm}$$

构造统计量：
数据总平均：

$$\bar{X} = \bar{X}_{\cdot\cdot} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}$$

因子水平 A_i 组内平均值：

$$\bar{X}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} X_{ij}$$

因子水平 B_i 组内平均值：

$$\bar{X}_{\cdot j} = \frac{1}{n}\sum_{i=n}^{m} X_{ij}$$

总离差平方和：

$$S_T = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X})^2$$

因子 A 效应平方和（组间）：

$$S_A = \sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{X}_{i\cdot} - \bar{X})^2$$

因子 B 效应平方和（组间）：

$$S_B = \sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{X}_{\cdot j} - \bar{X})^2$$

误差平方和：

$$S_e = \sum_{i=1}^{n}\sum_{j=1}^{m}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X})^2$$

同时有平方和分解公式：

$$S_T = S_A + S_B + S_e$$

由平方和分解定理：

$$S_T \sim \chi^2(nm-1),$$
$$S_A \sim \chi^2(n-1),\ S_B \sim \chi^2(m-1),\ S_e \sim \chi^2(n-1)(m-1)$$

于是可以构造两个检验统计量：

$$F_A = \frac{S_A/(n-1)}{S_e/(n-1)(m-1)}$$
$$F_B = \frac{S_B/(m-1)}{S_e/(n-1)(m-1)}$$

且满足：

$$F_A \sim F(n-1,(n-1)(m-1))$$
$$F_B \sim F(m-1,(n-1)(m-1))$$

否定域为：

$$F_A > F_{1-\alpha}(n-1,(n-1)(m-1))$$
$$F_B > F_{1-\alpha}(m-1,(n-1)(m-1))$$

平方和的计算公式

记

$$T_{\cdot\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} \quad T_{i\cdot} \sum_{j=1}^{m} X_{ij} \quad T_{\cdot j} = \sum_{i=1}^{n} X_{ij}$$

则可推知

$$S_A = \frac{1}{m}\sum_{i=1}^{n} T_{i\cdot}^2 - \frac{T_{\cdot\cdot}^2}{nm} \qquad S_B = \frac{1}{n}\sum_{j=1}^{n} T_{\cdot j}^2 - \frac{T_{\cdot\cdot}^2}{nm}$$
$$S_T = \sum i = 1^n \sum_{j=1}^{m} X_{ij}^2 - \frac{T_{\cdot\cdot}^2}{nm} \qquad S_e = S_T - S_A - S_B$$
$$TS \triangleq \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2$$

双因子间有相互作用的情形

假定 A, B 是影响实验结果的两个因子，因子 A 有 n 个水平：$A_1, A_2, \cdots, A_n$ ；因子 B 有 m 个水平：$B_1, B_2, \cdots, B_m$ 。在 A 的 n 个水平与 B 的 m 个水平的每种组合下分别作 r 次试验。 用 X_{ijk} 表示水平 A_iB_j 条件下第 k 次试验的结果，其观测值记为 x_{ijk} 。
NOTE: 为检验交互作用是否显著，各因素搭配下至少要做 2 次试验。否则交互作用与随机误差总是结合在一起，无法将二者分离开来。试验次数理论上越多越好，但受实际条件的限制，往往只能进行有限次，但是至少要大于两次。

假定在每种组合下试验结果 X_{ijk} 满足：

$$X_{ijk} \sim N(\mu_{ijk}, \sigma^2)$$

并设满足如下定义的效应可加性

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

其中 α_i 表示因子水平 A_i 的效应，表示因子 A 的第 i 个水平对试验结果带来的影响；β_j 为因子 B 的效应，表示因子 B 的第 j 个水平对试验结果带来的影响；γ_{ij} 为因子水平 A_i 和 B_j 的交互作用对试验结果带来的影响。

问题转化为检验如下假设：

$$H_0:\ \mu_{111} = \mu_{112} = \cdots = \mu_{ijk} = \cdots = \mu_{nmr}$$

构造统计量：
数据总平均：

$$\bar{X} = \bar{X}_{\cdots} = \frac{1}{nmr}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} X_{ijk}$$

因子水平 A_i 组内平均值：

$$\bar{X}_{i\cdot\cdot} = \frac{1}{mr}\sum_{j=1}^{m}\sum_{k=1}^{r} X_{ijk}$$

因子水平 B_i 组内平均值：

$$\bar{X}_{\cdot j\cdot} = \frac{1}{nr}\sum_{i=n}^{m}\sum_{k=1}^{r} X_{ijk}$$

总离差平方和：

$$S_T = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(X_{ijk} - \bar{X})^2$$

因子 A 效应平方和（组间）：

$$S_A = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(\bar{X}_{i\cdot\cdot} - \bar{X})^2$$

因子 B 效应平方和（组间）：

$$S_B = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(\bar{X}_{\cdot j\cdot} - \bar{X})^2$$

交互效应平方和：

$$S_{AB} = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2$$

误差平方和：

$$S_e = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(X_{ijk} - \bar{X}_{ij\cdot})^2$$

同时有平方和分解公式：

$$S_T = S_A + S_B + S_{AB} + S_e$$

由平方和分解定理：

$$\begin{gathered} S_T \sim \chi^2(nm-1), \\ S_A \sim \chi^2(n-1),\ S_B \sim \chi^2(m-1),\ S_{AB} \sim \chi^2(n-1)(m-1) \\ S_e \sim \chi^2(nm(r-1)) \end{gathered}$$

于是可以构造三个检验统计量：

$$\begin{aligned} F_A &= \frac{S_A/(n-1)}{S_e/nm(r-1)} \\ F_B &= \frac{S_B/(m-1)}{S_e/nm(r-1)} \\ F_{AB} &= \frac{S_{AB}/(n-1)(m-1)}{S_e/nm(r-1)} \end{aligned}$$

且满足：

$$\begin{aligned} F_A &\sim F(n-1, nm(r-1)) \\ F_B &\sim F(m-1, nm(r-1)) \\ F_{AB} &\sim F((n-1)(m-1), nm(r-1)) \end{aligned}$$

否定域为：

$$\begin{aligned} F_A &> F_{1-\alpha}(n-1, nm(r-1)) \\ F_B &> F_{1-\alpha}(m-1, nm(r-1)) \\ F_{AB} &> F_{1-\alpha}((n-1)(m-1), nm(r-1)) \end{aligned}$$