哈爾濱工業大學

第38讲 单因素试验的方差分析

方差分析

方差分析（**Analysis of Variance,简称ANOVA**）是基于样本方差的分解、分析鉴别一个变量或一些变量对一个特定变量的影响程度的统计分析方法。它是英国统计学家费歇尔在20世纪20年代创立的.

方差分析

方差分析用于推断多个正态总体在方差相同的条件下，均值是否相等的假设检验问题.

即，设有m个独立正态变量：

$$X_1 \sim N(\mu_1,\sigma^2), X_2 \sim N(\mu_2,\sigma^2),\cdots,$$

$$X_m \sim N(\mu_m,\sigma^2)$$

检验：$\mu_1=\mu_2=\cdots=\mu_m$ 是否成立

方差分析

例1 对m种不同的农药$A_1, A_2, \cdots, A_m$在相同条件下分别进行杀虫试验，设对农药A_i独立进行n_i次试验，结果如下. 问不同农药的杀虫率是否有显著差异？

农药 试验号	杀虫率				
	A_1	$\cdots$	A_i	$\cdots$	A_m
1	x_{11}	$\cdots$	x_{i1}	$\cdots$	x_{m1}
2	x_{12}	$\cdots$	x_{i2}	$\cdots$	x_{m2}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n_i	x_{1n_1}	$\cdots$	x_{in_i}	$\cdots$	x_{mn_m}

方差分析中的概念

在例1中，把试验结果“杀虫率”称为试验指标.

把影响试验结果的对象称为试验的因素，常用A, B等表示. 如例1中的“农药”即为因素.

只研究一个因素对试验指标影响的试验，称为单因素试验，否则称为多因素试验. 例1就是单因素试验.

单因素试验的方差分析

因素所取的不同的状态称为因素的水平. 常用$A_1, A_2 \cdots$表示. 如例1中的“m种不同的农药”即为因素“农药”的m个水平.

例1 属于单因素试验，解决例1问题的方法称为单因素试验的方差分析.

方差分析

例1 问不同农药的杀虫率是否有显著差异？

因素 → 农药　　杀虫率 ← 试验指标

试验号 \ 农药	杀虫率				
	A_1	$\cdots$	A_i	$\cdots$	A_m
1	x_{11}	$\cdots$	x_{i1}	$\cdots$	x_{m1}
2	x_{12}	$\cdots$	x_{i2}	$\cdots$	x_{m2}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n_i	x_{1n_1}	$\cdots$	x_{in_i}	$\cdots$	x_{mn_m}

- 纵向的差异称为随机误差(随机因素引起的)
- 横向的差异称为系统误差(农药不同引起的)

方差分析

假设农药A_i的杀虫率 $X_i \sim N(\mu_i, \sigma^2)(i=1,2,\cdots,m)$

且 $X_1, X_2, \cdots, X_m$ 独立.

在A_i下的试验数据 $X_{i1}, X_{i2}, \cdots, X_{in_i}$ 可看成

总体X_i容量为n_i的样本. 即

$$
\begin{array}{l}
A_1 : X_1 \sim N(\mu_1, \sigma^2),\ X_{11}, X_{12}, \cdots X_{1n_1} \\
A_2 : X_2 \sim N(\mu_2, \sigma^2),\ X_{21}, X_{22}, \cdots X_{2n_2} \\
\qquad \cdots \qquad\qquad \cdots \qquad\qquad \cdots \\
A_m : X_m \sim N(\mu_m, \sigma^2),\ X_{m1}, X_{m2}, \cdots X_{mn_m}
\end{array}
$$

问题：不同农药的杀虫率有无显著差异？

$\Longleftrightarrow$ 检验：$H_0: \mu_1 = \mu_2 \cdots = \mu_m$

$H_1: \mu_1, \mu_2, \cdots, \mu_m$不全相同

如何构造检验统计量呢？

单因素试验方差分析的数学模型

$$\begin{cases} X_{ij} = \mu_i + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0,\sigma^2) \end{cases} \quad (j=1,\cdots,n_i; i=1,\cdots m,)$$

其中μ_i 和σ^2是未知的参数，且各 ε_{ij}相互独立.

检验方法——偏差平方和分解

将引起数据总波动S的原因分解为

随机因素+因素不同水平A_i间的差异.

即，$S = S_e + S_A$

偏差平方和分解 $S = S_e + S_A$

总偏差平方和：$S = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2$ 数据总的波动

其中 $\bar{X} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}X_{ij}, \quad n = \sum_{i=1}^{m}n_i$.

误差平方和：$S_e = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2$

各组数据随机误差的累加引起的波动

组间平方和：$S_A = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot} - \bar{X})^2 = \sum_{i=1}^{m}n_i(\bar{X}_{i\cdot} - \bar{X})^2$

因素不同水平A_i间的差异引起的波动

方差分析

定理

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

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

(1) $\dfrac{S_e}{\sigma^2} \sim \chi^2(n-m)$;

(2) S_A与S_e相互独立;

(3) 当H_0为真时, $\dfrac{S_A}{\sigma^2} \sim \chi^2(m-1)$.

且统计量$F = \dfrac{S_A/(m-1)}{S_e/(n-m)} \sim F(m-1, n-m)$

方差分析

对给定的显著性水平α,

当$F > F_{\alpha}(m-1, n-m)$时，拒绝H_0.

单因素方差分析表

方差来源	平方和	自由度	均方	F值
因素A	S_A	$m-1$	$\bar{S}_A = S_A/(m-1)$	$\bar{S}_A/\bar{S}_e$
误差	S_e	$n-m$	$\bar{S}_e = S_e/(n-m)$	
总和	$S = S_A + S_e$	$n-1$		

方差分析

在单因素试验方差分析的数学模型中，

（1）未知参数μ_i的无偏估计量为

$$\hat{\mu}_i = \overline{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \quad (i = 1,2,\cdots,m)$$

（2）未知参数σ^2的无偏估计量为

$$\hat{\sigma}^2 = \frac{S_e}{n-m}$$

例1 对6种不同的农药在相同条件下分别进行杀虫试验，试验结果如下. 问不同农药的杀虫率是否有显著的差异？$(\alpha = 0.01)$

农药 \ 试验号	杀虫率					
	A_1	A_2	A_3	A_4	A_5	A_6
1	87	90	56	55	92	75
2	85	88	62	48	99	72
3	80	87			95	81
4		94			91	

设农药A_i的杀虫率 $X_i \sim N(\mu_i, \sigma^2)(i=1,\cdots,6)$

$H_0: \mu_1=\mu_2 \cdots = \mu_6$ $H_1: \mu_1, \mu_2, \cdots, \mu_6$不全相同

单因素方差分析表

方差来源	平方和	自由度	均方	F值
农药	3794.5	5	758.9	51.17
误差	178	12	14.83	
总和	3972.5	17		

查表得$F_{0.01}(5,12)=5.06$,由于$F=51.17>5.06$

故拒绝H_0, 说明农药杀虫率有显著差异.

谢　谢！