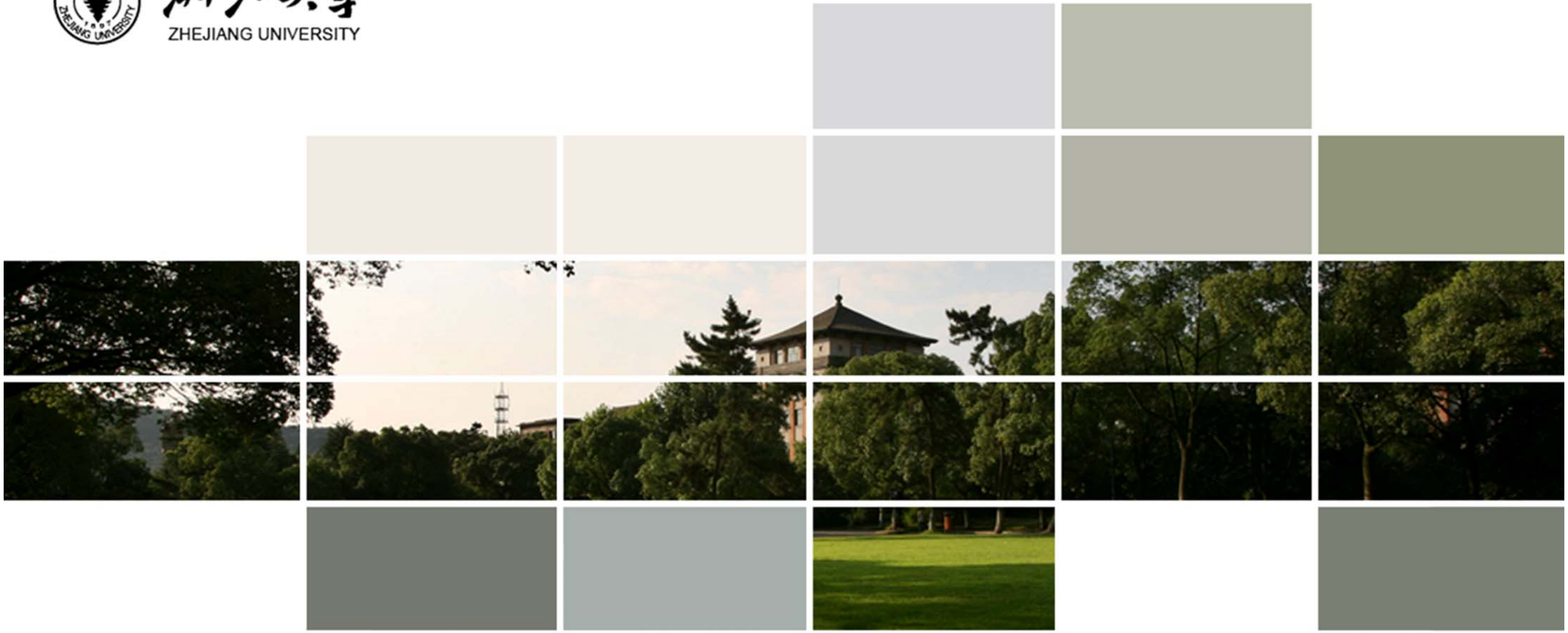
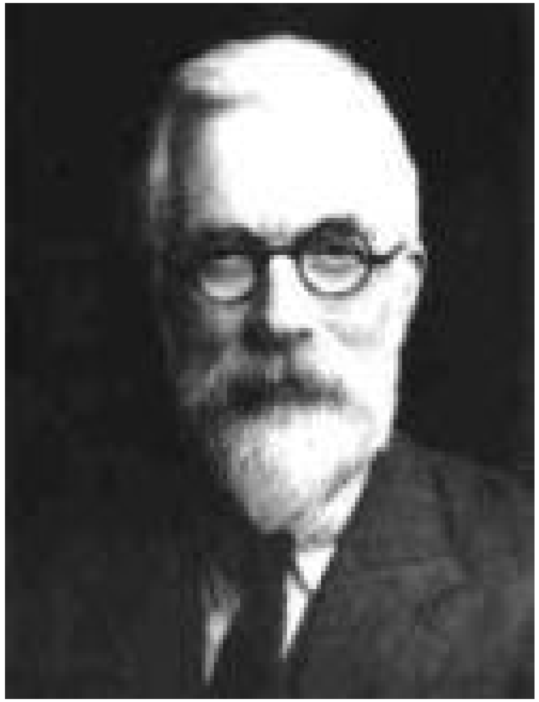
浙江大学
ZHEJIANG UNIVERSITY

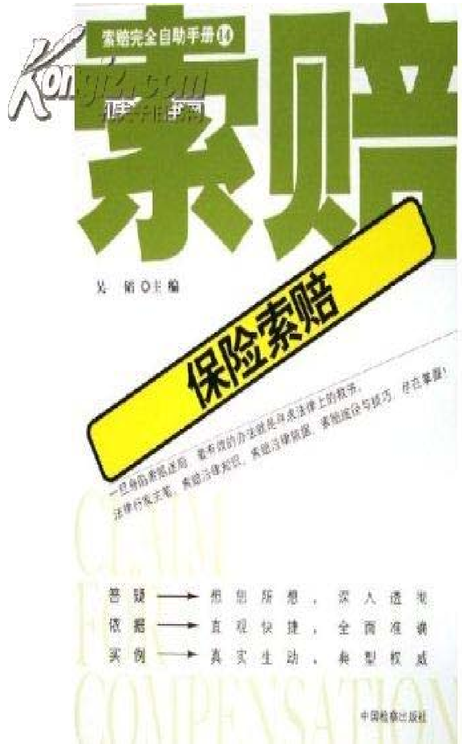


第61讲　单因素方差分析

- 方差分析(ANOVA),是由英国统计学家费歇尔(Fisher)在20世纪20年代提出的,可用于推断两个或两个以上总体均值是否有差异的显著性检验.



例1：保险公司为了解某一险种在四个不同地区索赔额情况是否存在差异。搜集了这四个不同地区一年的索赔额情况记录如表所示.　试判断在四个不同地区索赔额有无显著的差异?



保险索赔记录								
地区	索赔额(万元)							
A1	1.60	1.61	1.65	1.68	1.70	1.70	1.78	
A2	1.50	1.64	1.40	1.70	1.75			
A3	1.64	1.55	1.60	1.62	1.64	1.60	1.74	1.80
A4	1.51	1.52	1.53	1.57	1.64	1.60		

- 其一， 由于地区不同，而引起索赔额差异.
- 其二，同一地区，由于随机因素的影响， 也使其索赔额存在差异.

差异原因

- 在方差分析中，通常把研究对象的特征值，即所考察的试验结果称为试验指标.

本例中“索赔额”就是试验指标.

- 对试验指标产生影响的原因称为因素.

本例中“地区” 即为因素.

- 因素中各个不同状态称为水平.

本例中“四个不同地区”即为四个水平.

- 单因素方差分析 仅考虑有一个因素A对试验指标的影响. 假如因素A有r个水平, 分别在第 i 水平下进行了多次独立观测, 所得到的试验指标的数据.

$$
\boxed{\begin{array}{lcccc}
A_1 : N\left(\mu_1, \sigma^2\right) & X_{11} & X_{12} & \cdots & X_{1n_1} \\
A_2 : N\left(\mu_2, \sigma^2\right) & X_{21} & X_{22} & \cdots & X_{2n_2} \\
\vdots & \vdots & & \cdots & \vdots \\
A_r : N\left(\mu_r, \sigma^2\right) & X_{r1} & X_{r2} & \cdots & X_{rn_r}
\end{array}}
$$

各总体间相互独立. 因此，有如下的数学模型：

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

记$\sum_{i=1}^{r} n_i = n, \bar{X}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \bar{X} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij}.$

- 方差分析的目的就是要比较因素A 的r 个水平下试验指标理论均值的差异，问题可归结为比较这r个总体的均值差异.

检验假设 $H_0: \mu_1 = \mu_2 = \ldots = \mu_r$

$H_1: \mu_1, \mu_2, \ldots, \mu_r$不全相等。

- 检验假设采用的方法是平方和分解。即
- 假设数据总的差异用总离差平方和S_T表示，S_T分解为二个部分：
- 一部分是由于因素A引起的差异——效应平方和S_A；
- 另一部分则由随机误差所引起的差异——误差平方和S_E。

总偏差平方和 $S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2$

效应平方和 $S_A = \sum_{i=1}^{r} n_i\left(\bar{X}_{i\bullet} - \bar{X}\right)^2$

误差平方和 $S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_{i\bullet}\right)^2$

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)^2$$

$$S_A=\sum_{i=1}^{r}n_i\left(\bar{X}_{i\bullet}-\bar{X}\right)^2$$

$$S_E=\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{i\bullet}\right)^2$$

定理 (1) $S_T=S_A+S_E$，

(2) $\dfrac{S_E}{\sigma^2}\sim\chi^2(n-r)$;

(3) S_A 与 S_E 相互独立，

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

此时，$F=\dfrac{S_A/(r-1)}{S_E/(n-r)}\sim F(r-1,n-r)$.

单因素试验方差分析表

方差来源	平方和	自由度	均方	F
因素A	S_A	r-1	$MS_A=S_A/(r-1)$	MS_A/MS_E
误差	S_E	n-r	$MS_E=S_E/(n-r)$	
总和	S_T	n-1		

当$F \geq F_{\alpha}(r-1, n-r)$时，拒绝原假设.

方差分析通常都是直接用统计软件计算，在EXCEL中计算也非常方便。具体可以看实验25.

对例1的EXCEL结果如下：

表一：摘要

组	观测数	求和	平均	方差
行1	7	11.72	1.674	0.0038
行2	5	7.99	1.598	0.0210
行3	8	13.19	1.649	0.0067
行4	6	9.37	1.562	0.0026

表二：方差分析表

方差来源	平方和	自由度	均方	F	P-value	F crit
组间	0.0492	3	0.0164	2.1659	0.1208	3.0491
组内	0.1666	22	0.0076	$< F_{0.05}(3,22)$ 接受H_0.	> 0.05 接受H_0.	$F_{0.05}(3,22)$
总计	0.2158	25				