第4章 方差分析

方差分析 (***Analysis of Variance***，***ANOVA***) ：

研究一个(或多个)分类自变量如何影响一个数值因变量的统计分析方法。

方差分析针对方差相同的多个正态总体，检验它们的均值是否相同。 即，

同时判断多组数据均值之间差异是否显著

方差分析的目的

①. 判断某些因素对于我们感兴趣的因变量是否具有“显著”的影响，

②. 如果因素间有交互效应，寻找最佳搭配方案。

方差分析的特点

① 方差分析与一般的假设检验不同

要比较均值是否相同，可以使用第3 章假设检验的方法，但是只能处理两个均值。

方差分析处理的是多个均值的情况。

② 方差分析与回归、相关分析不同

回归与相关处理的是两个数值变量的问题，相应的散点在 x 轴上具有顺序(从小到大)，而方差分析的数据在 x 轴上可以任意交换位置。

常见的方差分析主要有：

单因素方差分析，双因素方差分析，

多因素方差分析。

4.1 方差分析的数学模型

响应变量(因变量)：

进行随机试验所考察的数值指标；

因素或因子(自变量)：

影响因变量的各不同分类原因；

水平：

各个因素所构成的组或者类型。

Fisher的农业试验

考察小麦产量(y) 对于品种和施肥量的关系。

选择了：两个不同的小麦品种，
三个不同的施肥等级；
一共 $2\times3=6$ 种搭配做试验，建立模型。

$$
\begin{cases}
y_{11} = \theta_0 + \alpha_1 + \beta_1 + \varepsilon_{11} \\
y_{12} = \theta_0 + \alpha_1 + \beta_2 + \varepsilon_{12} \\
y_{13} = \theta_0 + \alpha_1 + \beta_3 + \varepsilon_{13} \\
y_{21} = \theta_0 + \alpha_2 + \beta_1 + \varepsilon_{21} \\
y_{22} = \theta_0 + \alpha_2 + \beta_2 + \varepsilon_{22} \\
y_{23} = \theta_0 + \alpha_2 + \beta_3 + \varepsilon_{23}
\end{cases}
$$

y_{ij} 是小麦产量，
α_1、α_2 是品种效应，
β_1、β_2、β_3 是施肥等级的效应，
θ_0 是其它因素的平均效应。

ε_{ij} 是随机误差，$\text{i.i.d} \sim N(0, \sigma^2)$

品种是否对产量有影响 $\Leftrightarrow$ H_{01}：$\alpha_1 = \alpha_2$

施肥量是否对产量有影响 $\Leftrightarrow$ H_{02}：$\beta_1 = \beta_2 = \beta_3$

把这个模型写成矩阵的形式：$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{pmatrix}$$

在方差分析中，同一个因素的不同水平看成是模型里的不同变量，而不能看成是同一个自变量在不同试验里的取值。(否则需要 y 对 x 有线性相依关系)

4.2 单因素方差分析

1. 数据的结构

自变量水平	试验指标观察值				组内平均
1	y_{11}	y_{12}	…	y_{1n1}	$\bar{y}_1$
2	y_{21}	y_{22}	…	y_{2n2}	$\bar{y}_2$
	…	…		…	
r	y_{r1}	y_{r2}	…	$y_{r\,nr}$	$\bar{y}_r$

影响 y 的只有一个因素，它有 r 个水平(组)，在第 i 个水平下针对 y 做了 n_i 次试验或观察，得到因变量的观察数据为 y_{i1}，…， $y_{i\,ni}$。

可以假定：

$$y_{ij} = \beta_i + \varepsilon_{ij} \quad ，\ 1 \le j \le n_i \ 、\ 1 \le i \le r$$

这里 ε_{ij} 对所有 i、j 都独立同分布于 $N(0, \sigma^2)$

单因素方差分析的主要任务：

1. 检验假设：H_0： $\beta_1 = \beta_2 = \ldots = \beta_r$ ；
2. 作出未知参数 β_1，…， β_r 以及 σ^2 的估计

2. 因子效应与误差方差的估计

按照模型的假定，因变量的观察值来自 r 个不同的正态总体：

$y_{11}, \cdots, y_{1n1}$ 来自总体 $N(\beta_1, \sigma^2)$；

$y_{21}, \cdots, y_{2n2}$ 来自总体 $N(\beta_2, \sigma^2)$；

…

$y_{r1}, \cdots, y_{rnr}$ 来自总体 $N(\beta_r, \sigma^2)$。

未知参数 $\beta_1, \cdots, \beta_r$ 的估计就采用各个总体的样本均值。

定理 4.1 方差分析中未知参数估计及分布

1. 因素各水平效应的估计采用各个组内平均，

$$\hat{\beta}_i = \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

相应的分布显然是：$\hat{\beta}_i \sim N(\beta_i, \frac{\sigma^2}{n_i})$ $1 \le i \le r$

2. 误差方差 σ^2 的估计利用残差平方和，

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-r} = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_s}(y_{ij} - \bar{y}_i)^2$$

相应的分布是：$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-r)$

3. $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_r, \hat{\sigma}^2$ 之间相互独立。

3. 方差分析平方和分解公式

(1) 总平方和 TSS $= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$

它是观察到的每个数据与总平均的差异总和，表示因变量总的变化。

TSS衡量了全部 y_{ij} 的差异，它越大则说明y_{ij}之间的差异越大。产生TSS的原因只有两个：

(1). 因子不同的水平，即 β_1 ，…， β_r 的差异；
(2). 随机误差。

(2) 自变量平方和 CSS $= \sum_{i=1}^{r} n_i (\bar{y}_i - \bar{y})^2$

它是因为自变量不同的类型而产生的差异，表示自变量在因变量的变化中所占的份额。

Σ (每组平均 - 总平均)^2 来刻化

(3) 残差平方和 RSS $= \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$

表示由其它原因引起的因变量变化，利用

Σ (观察值- 每组平均)^2 来刻化

方差分析平方和分解

$$TSS = CSS + RSS$$

利用Cochren定理可以证明，

自变量平方和 CSS 与残差平方和 RSS 相互独立。

4. 单因素方差分析的检验

如果零假设H_0：$\beta_1 = \beta_2 = \ldots = \beta_r$ 成立，则有

$$\frac{CSS}{\sigma^2} \sim \chi^2(r-1)$$

由定理4.1.1 构造检验统计量

$$F\text{ 比}= \frac{n-r}{r-1}\frac{CSS}{RSS} \sim F(r-1, n-r)$$

因此这些分类自变量中每组均值都相同的一个水平α的拒绝域为：$\{F \geq F_\alpha(r-1, n-r)\}$

检验的p-值是 $P\{F(r-1, n-r) > F\text{ 比}\}$

单因素方差分析表

方差来源	平方和	自由度	均方	F-比	p-值
分类变量	CSS	r-1	CMS		
残差变量	RSS	n-r	RMS		
总计	TSS	n-1			

其中

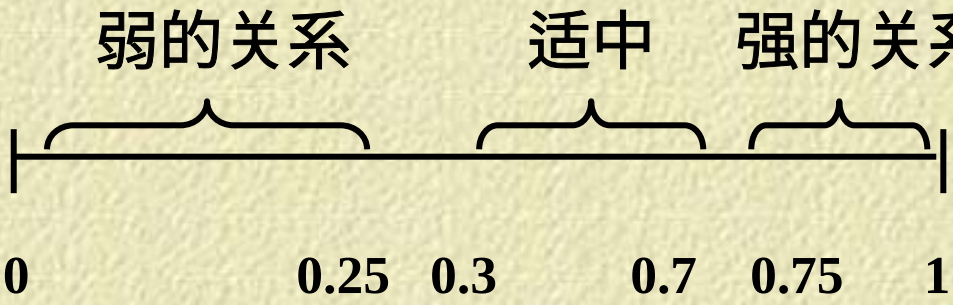
$$\text{CMS} = \frac{\text{CSS}}{r\text{-}1},\ \text{RMS} = \frac{\text{RSS}}{n\text{-}r},\ F\text{-比} = \frac{\text{CMS}}{\text{RMS}}$$

5. 变量关系的强度

$$R^2 = \frac{\text{自变量平方和}}{\text{总平方和}} = \frac{\text{CSS}}{\text{TSS}}$$

R^2 反映了在因变量全部的变化中，分类自变量产生影响所占的比例；

因此用 R (取正值)来衡量分类自变量与数值因变量的关系强度：



EXCEL 函数处理单因素方差分析

直接调用函数 ***DEVSQ***

计算出总平方和以及残差平方和。

$$DEVSQ(x_1,\cdots,x_n)=\sum_{k=1}^{n}(x_k-\bar{x})^2$$

(1) 没有必要计算总均值以及各组均值；

(2) 先计算各组偏差平方和 RS1, …, RSr；
再全部相加得到残差平方和 RSS。

例4.1 灯丝配料方案的优选

灯丝	使用寿命(小时)
甲	1600，1610，1650，1680，1700，1720，1800；
乙	1580，1640，1640，1700，1750；
丙	1460，1550，1600，1640，1660，1740， 1820， 1820；
丁	1510，1520，1530，1570，1600，1680；

解. $n = 26$，$r = 4$。

方差来源	平方和	自由度	均方	F-比	p-值
分类变量	47399.17	3	15799.7	1.9327	0.1538
残差变量	179850.8	22	8175.04		
总计	227250	25			

水平 0.05 下认为灯丝配料对寿命没有显著影响。

思考1

这里的R^2 =0.209 应该如何理解？

□

例4.2

下表是 *FBI* 给出的1986~1992年美国 48 个大陆洲暴力犯罪率(次/10 万人)，按地理位置分成 7 个区。

新英格兰	147 140 149 557 336 426
中大西洋	986 572 359
中西部	423 308 800 804 258 285
	235 263 578 51 125 369
南方	427 833 306 164 476 675 588
	1036 334 540 558 274 395 758
西南	436 659 658 726
落基山区	293 524 157 222 267 719
太平洋岸	437 550 920

作方差分析判断犯罪率是否与地区有关？

解.

***p*-值是 0.07224，因此在水平0.05下不能拒绝零假设，即应该认为在统计上没有显著的证据表明犯罪率和地区有关(各地犯罪率没有显著差异)。**

R^2 **=0.236 说明在影响犯罪率的所有因素中，地理位置占了将近四分之一的比例。**

□

例4.3 2003年9月全国城镇居民家庭总收入

说明：去掉无数据的西藏等地，把其余30个省市按地理位置划分为六个大区：

东北：辽宁、吉林、黑龙江；
华北：北京、天津、河北、山西、内蒙；
西北：陕西、甘肃、宁夏、青海、新疆；
东南：上海、江苏、浙江、安徽、福建、
　　江西、山东、河南、湖北、湖南；
华南：广东、广西、海南；
西南：重庆、四川、云南、贵州；

东北：671，606，600；
华北：1234，901，650，617，654；
西北：643，601，586，580，686；
东南：1379，835，1137，634，884，634，
766，612，650，678；
华南：1102，671，640；
西南：738，639，576，695

我们需要通过统计分析，讨论在2003年9月份我国大陆地区城镇居民的收入是否依地区的不同而有显著差异？

p-值是0.37，不能拒绝零假设，因此可以认为这些地区的居民收入没有显著差异，即2003年9月我国城镇经济发展水平比较均衡。

$R^2 = 0.19$，说明城镇居民的收入有 1/5 与地理位置有关。

思考2

是否应该把北京、上海等偏高的样本排除？

思考3

还有没有其它更好的分类方法？ □

4.3 双因素方差分析

假定影响 y 的有两个因素A、B，各有r、 s 个水平：$A_1, \ldots, A_r$, $B_1, \ldots, B_s$, 对这些水平所有的搭配 $A_i B_j$ 同时都做 l 次($l>1$)试验得到模型：

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

$$1 \le i \le r, 1 \le j \le s, 1 \le k \le l, \varepsilon_{ijk} \text{ i.i.d} \sim N(0, \sigma^2)$$

这里 μ 是响应变量 y 的总平均；

α_i :A 的主效应，A 在第i 个水平单独对 y 的效果；

β_j :B 的主效应，B 在第j 个水平单独对 y 的效果；

γ_{ij} :交互效应，因素A 在 i 、因素B 在 j 水平上联合对 y 的效果。

一般为了便于分析，还做如下假定：

$$\sum_{i=1}^{r}\alpha_i=\sum_{j=1}^{s}\beta_j=0=\sum_{i=1}^{r}\gamma_{ij}=\sum_{j=1}^{s}\gamma_{ij}$$

双因素方差分析需要讨论：

1. 因子的主效应是否显著；即检验：
   H_{01}：$\alpha_1=\alpha_2=\ldots=\alpha_r$，以及$H_{02}$：$\beta_1=\beta_2=\ldots=\beta_s$
2. 交互效应是否显著：H_{03}：$\gamma_{11}=\gamma_{12}=\ldots=\gamma_{rs}$

与单因素方差分析不同的是，如果拒绝了 H_{03}，还应该寻找最佳搭配。

处理思路类似单因素方差分析，把总平方和分解成若干平方和：有两个因素主效应产生的，有交互效应产生的，还有随机误差产生的，最后构造恰当的 F 统计量来检验 H_{01}、H_{02}、H_{03}。

引进如下符号：

总平均：

$$\bar{y} = \frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{l} y_{ijk}$$

误差平均：

$$\bar{y}_{ij\bullet} = \frac{1}{l}\sum_{k=1}^{l} y_{ijk}$$

*A*因素平均：

$$\bar{y}_{i\bullet\bullet} = \frac{1}{s}\sum_{j=1}^{s} \bar{y}_{ij\bullet}$$

*B*因素平均：

$$\bar{y}_{\bullet j\bullet} = \frac{1}{r}\sum_{i=1}^{r} \bar{y}_{ij\bullet}$$

构造如下相应的五个平方和：

总平方和 $$\mathrm{TSS}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{l}(y_{ijk}-\bar{y})^2$$

*A*因子主效应平方和 $$\mathrm{SSA}=sl\sum_{i=1}^{r}(\bar{y}_{i\cdot\cdot}-\bar{y})^2$$

*B*因子主效应平方和 $$\mathrm{SSB}=rl\sum_{j=1}^{s}(\bar{y}_{\cdot j\cdot}-\bar{y})^2$$

交互效应平方和 $$\mathrm{SSAB}=l\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y})^2$$

随机误差平方和 $$\mathrm{RSS}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{l}(y_{ijk}-\bar{y}_{ij\cdot})^2$$

仍然可以利用Cochren 定理证明，

$$TSS = SSA + SSB + SSAB + RSS$$

而且等式右边的四个平方和相互独立

(1) $\frac{RSS}{\sigma^2} \sim \chi^2(rs(l-1))$ ；

(2) 当H_{01} 成立时，$\frac{SSA}{\sigma^2} \sim \chi^2(r-1)$ ；

(3) 当H_{02} 成立时，$\frac{SSB}{\sigma^2} \sim \chi^2(s-1)$ ；

(4) 当H_{03} 成立时，$\frac{SSAB}{\sigma^2} \sim \chi^2((r-1)(s-1))$ 。

因此可以构造出三个零假设的 F 检验，

1. 对于零假设 H_{01}：$\alpha_1=\alpha_2=\ldots=\alpha_r$，构造

检验统计量 $F_A = \dfrac{rs(l-1)}{r-1}\dfrac{SSA}{RSS} \sim F(r-1, rs(l-1))$

双因素方差分析中A因素的主效应不显著的一个检验水平α的拒绝域为：

$$\{ F_A \geq F_\alpha(r-1, rs(l-1)) \}$$

2. 对于零假设 H_{02}：$\beta_1=\beta_2=\ldots=\beta_s$，构造

检验统计量 $F_B = \dfrac{rs(l-1)}{s-1}\dfrac{SSB}{RSS} \sim F(s-1, rs(l-1))$

双因素方差分析中B因素的主效应不显著的一个检验水平α的拒绝域为：

$$\{ F_B \geq F_\alpha(s-1, rs(l-1)) \}$$

3. 对于零假设 H_{03}：$\gamma_{11} = \gamma_{12} = \ldots = \gamma_{rs}$，构造检验统计量 F_{AB}：

$$F_{AB} = \frac{rs(l-1)}{(r-1)(s-1)} \frac{SSAB}{RSS} \sim F((r-1)(s-1), rs(l-1))$$

双因素方差分析中A、B两个因素的交互效应不显著的一个检验水平α拒绝域为：

$$\{F_{AB} \geq F_{\alpha}((r-1)(s-1), rs(l-1))\}$$

双因素方差分析表

方差来源	平方和	自由度	平均平方和	F 值
因素A	SSA	$r-1$	$\frac{SSA}{r-1}$	$\frac{rs(l-1)}{r-1}\frac{SSA}{RSS}$
因素B	SSB	$s-1$	$\frac{SSB}{s-1}$	$\frac{rs(l-1)}{s-1}\frac{SSB}{RSS}$
$A\times B$	SSAB	$(r-1)(s-1)$	$\frac{SSAB}{(r-1)(s-1)}$	$\frac{rs(l-1)}{(r-1)(s-1)}\frac{SSAB}{RSS}$
残差	RSS	$rs(l-1)$	$\frac{RSS}{rs(l-1)}$	
总和	TSS	$rsl-1$		

EXCEL 函数处理双因素方差分析

① TSS = *DEVSQ*(全部试验数据)；

② *A* 因子主效应 SSA 的计算：

把 *A* 因子每个水平所在组的数据相加，得到 *r* 个和，计算DEVSQ(*A*的每组数据之和)，

$$SSA = \frac{DEVSQ(A\text{的每组数据之和})}{s\ l}$$

③ B 因子主效应 SSB 的计算：

把 B 因子每个水平所在组的数据相加，得到 s 个和，计算DEVSQ(B的每组数据之和)，

$$SSB = \frac{DEVSQ(B\text{的每组数据之和})}{r\,l}$$

④ 残差平方和 RSS 的计算：

计算每一对搭配 $i \times j$ 的 l 个重复试验数据的偏差平方和，再对所有这些 $r\,s$ 个 $DEVSQ$ 求和得到 RSS。

例4.4　橡胶配方中考虑3 种促进剂(A) 、4种氧化锌份量(B)各组合两次进行试验，测得24 组300%的定伸强力数据，对这组数据作双因素方差分析。

	B_1	B_2	B_3	B_4
A_1	31,33	34,36	35,36	39,38
A_2	33,34	36,37	37,39	38,41
A_3	35,37	37,38	39,40	42,44

解. 根据模型的定义，$r = 3$，$s = 4$，$l = 2$

方差来源	平方和	自由度	均方	F-值
A(促进剂)	56.59	2	28.29	$F_A = 19.4$
B(氧化锌)	132.125	3	44.04	$F_B = 30.2$
交互 $A \times B$	4.75	6	0.7917	$F_{AB} = 0.5429$
误差	17.5	12	1.4583	
总和	210.9583	23		

相应的 F 分布上0.05 分位点是：

$F_{0.05}(3,12) = 3.49$, $F_{0.05}(2,12) = 3.89$, $F_{0.05}(6,12) = 3.00$

所以在0.05的显著水平下，H_{01}、 H_{02}被否定，即不同的促进剂或不同的氧化锌分量对橡胶定伸强力具有显著的影响；

但是接受H_{03}，即交互作用不显著的。

从*p*-值的角度，

H_{01} 的 *p*-值是 0.000174 ；

H_{02} 的 *p*-值是 0.000007 ；

H_{03} 的 *p*-值有 0.766517。 □

例4.5

对一种火箭使用4 种燃料，3 种推进器进行射程试验，每种燃料与推进器各组合两次，一共试验了 24 次。

(单位：海里)

燃料	推进器 B_1	B_2	B_3
A_1	58.2	56.2	65.3
	52.6	41.2	60.8
A_2	49.1	54.1	51.6
	42.8	50.5	48.4
A_3	60.1	70.9	39.2
	58.3	73.2	40.7
A_4	75.8	58.2	48.7
	71.5	51.0	41.4

对这组数据作双因素方差分析。

解. 根据模型的定义，$r=4$，$s=3$，$l=2$

方差来源	平方和	自由度	平均平方和	F 值
A(燃料)	261.6750	3	87.2250	$F_A = 4.42$
B(推进器)	370.9808	2	185.4904	$F_B = 9.39$
交互 $A\times B$	1768.6925	6	294.7821	$F_{AB} = 14.9$
误差	236.95	12	19.7458	
总和	2638.2983	23		

相应的 F 分布上0.05 分位点是：

$F_{0.05}(3,12)=3.49, F_{0.05}(2,12)=3.89, F_{0.05}(6,12)=3.00$

所以在0.05的显著水平下，这三个零假设 H_{01}、H_{02}、H_{03} 都被否定。即，不同的燃料或不同的推进器对射程有统计上的显著影响；而且交互作用是高度显著的($A_4 \times B_1$ 或 $A_3 \times B_2$)

从*p*-值的角度，

H_{01} 的 *p*-值是 0.025969 ；

H_{02} 的 *p*-值是 0.003506 ；

H_{03} 的 *p*-值只有 0.000062。 □

思考4

在双因素方差分析中，“$l>1$”即需要做重复试验的目的是什么？什么情况下可以不做重复试验？

练习4.6

讨论没有重复试验的双因素方差分析模型。

练习4.7

如何用 *EXCEL* 处理没有重复试验的双因素方差分析？