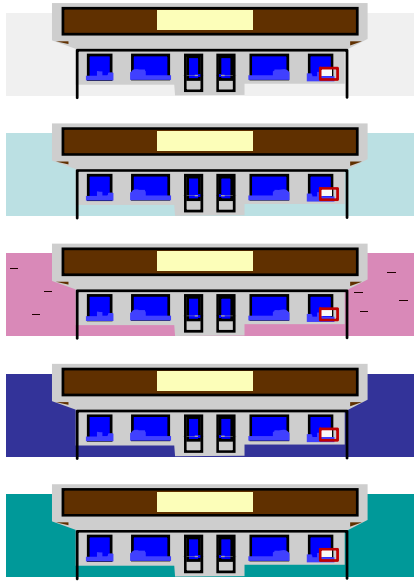
第七章　方差分析

什么是方差分析?

【例.1】某饮料生产企业研制出一种新型饮料。饮料的颜色共有四种，分别为橘黄色、粉色、绿色和无色透明。这四种饮料的营养含量、味道、价格、包装等可能影响销售量的因素全部相同。现从地理位置相似、经营规模相仿的五家超级市场上收集了前一时期该饮料的销售情况，见表1。试分析饮料的颜色是否对销售量产生影响。

表1　该饮料在五家超市的销售情况

超市	无色	粉色	橘黄色	绿色
1	26.5	31.2	27.9	30.8
2	28.7	28.3	25.1	29.6
3	25.1	30.8	28.5	32.4
4	29.1	27.9	24.2	31.7
5	27.2	29.6	26.5	32.8



什么是方差分析?

1. 检验饮料的颜色对销售量是否有影响，也就是检验四种颜色饮料的平均销售量是否相同
2. 设μ_1为无色饮料的平均销售量，μ_2粉色饮料的平均销售量，μ_3为橘黄色饮料的平均销售量，μ_4为绿色饮料的平均销售量，也就是检验下面的假设
   - H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
   - H_1: μ_1 , μ_2 , μ_3 , μ_4 不全相等
3. 检验上述假设所采用的方法就是方差分析 (ANOVA: analysis of variance)

什么是方差分析?

1. 检验多个总体均值是否相等
   - 通过对各观察数据误差来源的分析来判断多个总体均值是否相等
2. 变量

   一个定类尺度的自变量（k 个处理水平或分类）

   一个定距或比例尺度的因变量
3. 用于分析完全随机化试验设计

因素与水平

1、因素与水平

- 因素（factor）：影响试验指标的原因，用A、B、C等字母表示;
- 水平（level）：试验因素的不同状态，表示为A_1、A_2、B_1、B_2。
- 观察值：在每个因素水平下得到的样本值

2、可控因素与非控因素

- 可控因素或固定因素：因素的水平可准确控制
- 非控因素或随机因素：因素的水平不能严格控制，或效应不完全由因素水平控制

处理与重复

- 处理（treatment）
  - 对受试对象给予的某种外部干预或措施
  - 单因素处理与多因素处理

- 重复（repetition）
  - 将一个处理实施在两个或多个试验单位上

第一节　方差分析的基本问题

一. 方差分析的原理

二. 方差分析的基本假定

方差分析的基本思想和原理

方差分析的基本思想和原理

1. 随机误差

- 在因素的同一水平(同一个总体)下，样本的各观察值之间的差异
- 比如，同一种颜色的饮料在不同超市上的销售量是不同的
- 不同超市销售量的差异可以看成是随机因素的影响，或者说是由于抽样的随机性所造成的，称为***随机误差***

2. 系统误差

- 在因素的不同水平(不同总体)下，各观察值之间的差异
- 比如，同一家超市，不同颜色饮料的销售量也是不同的
- 这种差异*可能*是由于抽样的随机性所造成的，*也可能*是由于颜色本身所造成的，后者所形成的误差是由系统性因素造成的，称为***系统误差***

方差分析的基本思想和原理
（两类方差）

1. **组内方差**
   - 因素的同一水平(同一个总体)下样本数据的方差
   - 比如，无色饮料A_1在5家超市销售数量的方差
   - 组内方差只包含*随机误差*
2. **组间方差**
   - 因素的不同水平(不同总体)下各样本之间的方差
   - 比如，A_1、A_2、A_3、A_4四种颜色饮料销售量之间的方差
   - 组间方差既包括*随机误差*，也包括*系统误差*

方差分析的基本思想和原理

（方差的比较）

1. 如果不同颜色(水平)对销售量(结果)没有影响，那么在组间方差中只包含有随机误差，而没有系统误差。这时，组间方差与组内方差就应该很接近，两个方差的比值就会接近1
2. 如果不同的水平对结果有影响，在组间方差中除了包含随机误差外，还会包含有系统误差，这时组间方差就会大于组内方差，组间方差与组内方差的比值就会大于1
3. 当这个比值大到某种程度时，就可以说不同水平之间存在着显著差异

方差分析中的基本假定

方差分析中的基本假定

1. 每个总体都应服从正态分布
   - 对于因素的每一个水平，其观察值是来自服从正态分布总体的简单随机样本
   - 比如，每种颜色饮料的销售量必需服从正态分布
2. 各个总体的方差必须相同
   - 对于各组观察数据，是从具有相同方差的总体中抽取的
   - 比如，四种颜色饮料的销售量的方差都相同
3. 观察值是独立的
   - 比如，每个超市的销售量都与其他超市的销售量独立

方差分析中基本假定

- 如果原假设成立，即H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
  - 四种颜色饮料销售的均值都相等
  - 没有系统误差

  这意味着每个样本都来自均值为 μ、差为σ^2的同一正态总体

- 如果备择假设成立，即H_1: μ_i (i=1，2，3，4)不全相等
  - 至少有一个总体的均值是不同的
  - 有系统误差

  这意味着四个样本分别来自均值不同的四个正态总体

第二节　单因素方差分析

一. 方差分析的数学模型

二. 单因素方差分析的步骤

三. 方差分析中的多重比较

四. 单因素方差分析中的其他问题

数学模型

单因素方差分析的数据结构

重复（j）	因素(A) i			
	水平A_1	水平A_2	…	水平A_k
1	x_{11}	x_{12}	…	x_{1k}
2	x_{21}	x_{22}	…	x_{2k}
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{nk}

线性统计模型： $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

式中：μ 为总平均数

α_i 为第 i 水平的处理效应

ε_{ij}为随机误差，$\varepsilon_{ij} \sim N(0, \sigma^2)$

一、固定效应模型

$$\sum \alpha_i = 0$$

二、随机效应模型

$$\alpha_i \sim N(0, \sigma^2)$$

平方和与自由度的分解

（一）平方和分解

$$x_{ij} - \bar{\bar{x}} = \left(x_{ij} - \bar{x}_i\right) + \left(\bar{x}_i - \bar{\bar{x}}\right)$$

其中，

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \qquad (i = 1,2,\Lambda\ ,k)$$

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n}$$

式中：$n = n_1 + n_2 + \Lambda\ + n_k$

$$(x_{ij}-\bar{\bar{x}})^2 = (x_{ij}-\bar{x}_i)^2 + 2\left(x_{ij}-\bar{x}_i\right)(\bar{x}_i-\bar{\bar{x}}) + (\bar{x}_i-\bar{\bar{x}})^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{\bar{x}})^2$$

$$+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)(\bar{x}_i-\bar{\bar{x}})$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + \sum_{i=1}^{k} n_i(\bar{x}_i-\bar{\bar{x}})^2$$

$$+2\sum_{i=1}^{k}\left[(\bar{x}_i-\bar{\bar{x}})\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)\right]$$

- 总离差平方和(*SST*)、误差项离差平方和(*SSE*)、水平项离差平方和 (*SSA*) 之间的关系

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{\bar{x}})^2$$

$$SST = SSE + SSA$$

1、SST 反映全部观察值的离散状况；

2、SSE 反映每个总体各观察值的离散状况，又称组内离差平方和，反映的是随机误差的大小；

3、SSA 反映各总体的样本均值之间的差异程度，又称组间平方和，既包括随机误差也包括系统误差。

平方和与自由度的分解

（二）自由度分解

$$df_T = df_E + df_A$$

- ***SST***的自由度df_T为*n*-1，其中n为全部观察值的个数
- ***SSA***的自由度df_A为*k*-1，其中k为因素水平(总体)的个数
- ***SSE*** 的自由度df_E为*n-k*

平方和与自由度的分解

（三）计算方差

1. SSA的均方也称组间方差，记为MSA，计算公式为

$$MSA = \frac{SSA}{k-1}$$

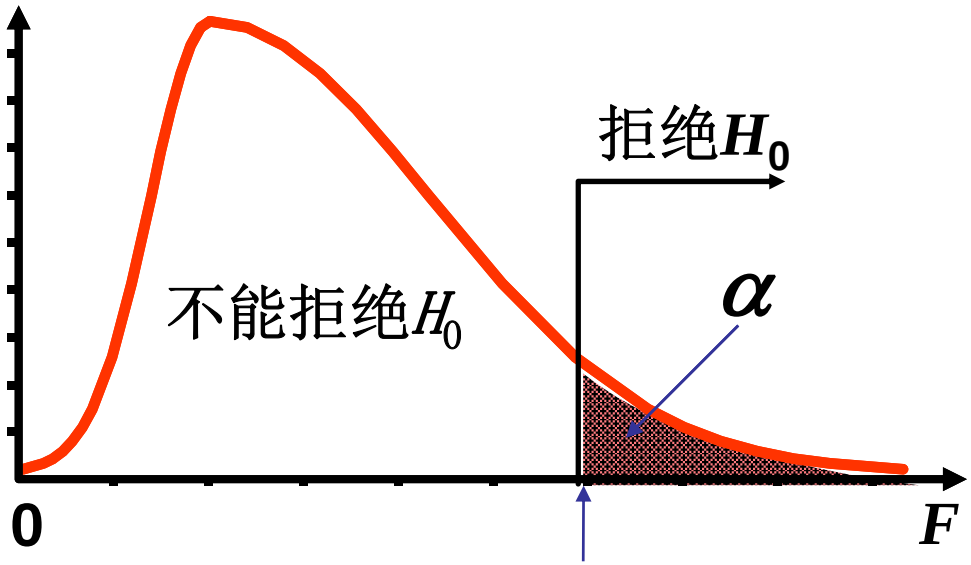
2. SSE的均方也称组内方差，记为MSE，计算公式为

$$MSE = \frac{SSE}{n-k}$$

构造检验的统计量

$$F = \frac{MSA}{MSE} \sim F(k-1, n-k)$$

如果均值相等，
$F=MSA/MSE \to 1$



F 分布

单因素方差分析的步骤

- 提出假设
- 构造检验统计量
- 统计决策

提出假设

1. 一般提法
   - H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (因素有k个水平）
   - H_1: μ_1，μ_2，…，μ_k不全相等
2. 对前面的例子
   - H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
     - 颜色对销售量没有影响
   - H_0: μ_1，μ_2，μ_3，μ_4不全相等
     - 颜色对销售量有影响

构造检验的统计量

表2　四种颜色饮料的销售量及均值					
超市（j）	水平A（i）				
	无色(A_1)	粉色(A_2)	橘黄色(A_3)	绿色(A_4)	
1	26.5	31.2	27.9	30.8	
2	28.7	28.3	25.1	29.6	
3	25.1	30.8	28.5	32.4	
4	29.1	27.9	24.2	31.7	
5	27.2	29.6	26.5	32.8	
合计	136.6	147.8	132.2	157.3	573.9
水平均值	$\bar{x}_1$ =27.32	$\bar{x}_2$=29.56	$\bar{x}_3$=26.44	$\bar{x}_4$=31.46	总均值
观察值个数	n_1=5	n_2=5	n_3=5	n_4=5	$\bar{\bar{x}}$ =28.695

单因素方差分析表
(基本结构)

方差来源	平方和 *SS*	自由度 *df*	均方 *MS*	*F* 值
组间(因素影响)	***SSA***	***k*-1**	***MSA***	$\frac{MSA}{MSE}$
组内(误差)	***SSE***	***n*-*k***	***MSE***	
总和	***SST***	***n*-1**		

单因素方差分析

方差分析：单因素方差分析

SUMMARY

组	计数	求和	平均	方差
列 1	5	136.6	27.32	2.672
列 2	5	147.8	29.56	2.143
列 3	5	132.2	26.44	3.298
列 4	5	157.3	31.46	1.658

方差分析

差异源	SS	df	MS	F	P-value	F crit
组间	76.8455	3	25.615	10.486	0.00047	3.2389
组内	39.084	16	2.4428			

统计决策

- 将统计量的值F与给定的显著性水平α的临界值F_α进行比较，作出接受或拒绝原假设H_0的决策
  - 根据给定的显著性水平α，在F分布表中查找与第一自由度$df_1 = k\text{-}1$、第二自由度$df_2 = n\text{-}k$ 相应的临界值 F_α
  - 若$F > F_\alpha$，则拒绝原假设H_0，表明均值之间的差异是显著的，所检验的因素(A)对观察值有显著影响
  - 若$F \le F_\alpha$，则不能拒绝原假设H_0，表明所检验的因素(A)对观察值没有显著影响

平方和的简易计算

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2 - n\bar{\bar{x}}^2$$

$$SSA = \sum_{i=1}^{k}n_i(\bar{x}_i - \bar{\bar{x}})^2 = \sum_{i=1}^{k}n_i\bar{x}_i^2 - n\bar{\bar{x}}^2$$

令 $C = n\bar{\bar{x}}^2 = \dfrac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}\right)^2}{n}$

$$SSE = SST - SSA$$

单因素方差分析

【例】某花卉研究所为促进芦荟生长，提高其经济效益，研究了4种不同配方的营养土对芦荟生长的影响。选取初始高度一致的试管苗20株，随机分成4组，一段时间后测量各处理试管苗株高，数据如表所示，试进行方差分析。($\alpha=0.05$)

单因素方差分析

四种不同营养土培养下芦荟的株高					
观察值	营养土(A)				
(j)	A_1	A_2	A_3	A_4	
1	**18.1**	**17.4**	**17.3**	**15.6**	
2	**18.6**	**17.9**	**16.9**	**15.8**	
3	**18.7**	**17.1**	**18.5**	**16.7**	
4	**18.9**	**16.5**	**18.2**	**15.3**	
5	**18.3**	**17.5**	**16.2**	**16.8**	
合计	**92.6**	**86.4**	**87.1**	**80.2**	$\sum$=346.3
平均数	**18.52**	**17.28**	**17.42**	**16.04**	$\bar{\bar{x}}$ =17.32

解：设四种营养土培养后的苗株高的均值分别为μ_1、μ_2、μ_3、μ_4，则需要检验如下假设

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (四种营养土效果无显著差异)

H_1: μ_1、μ_2、μ_3、μ_4不全相等 (有显著差异)

$$C = n\bar{\bar{x}}^2 = 5996.18$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - C = 22.31$$

$$SSA = \sum_{i=1}^{k} n_i \bar{x}_i^2 - C = 15.45$$

$$SSE = SST - SSA = 6.86$$

$$F = \frac{MSA}{MSE} = \frac{SSA/df_A}{SSE/df_E}$$

$$= \frac{15.45/3}{6.86/16} = 11.98$$

$$F_{0.05}(3, 16) = 3.24$$

- 结论：拒绝H_0。不同营养土培养下芦荟株高的差异非常显著。

方差分析表

差异源	SS	自由度	MS	F	临界值
组间	15.45	3	5.15	11.98	3.24
组内	6.86	19	0.43		
总和	22.31	22			

- 例：为了探讨不同窝的动物出生重是否存在差异，随机选取4窝动物，每窝中均有4只幼仔，结果见表：

动物编号	窝别			
	I	II	III	IV
1	34.7	33.2	27.1	32.9
2	33.3	26.0	23.3	31.4
3	26.2	28.6	27.8	25.7
4	31.6	32.3	26.7	28.0
和	125.8	120.1	104.9	118.0
均值	31.450	30.025	26.225	29.500

解：H_0: $\sigma^2 = 0$ (不同窝动物出生重无显著差异)

H_1: $\sigma^2 > 0$ (有显著差异)

$$C = n\bar{\bar{x}}^2 = 13735.84$$

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - C = 177.52$$

$$SSA = \sum_{i=1}^{k} n_i \bar{x}_i^2 - C = 58.575$$

$$SSE = SST - SSA = 118.945$$

$$F = \frac{MSA}{MSE} = \frac{SSA/df_A}{SSE/df_E}$$

$$= \frac{58.575/3}{118.945/12} = 1.97$$

$$F_{0.05}(3, 12) = 3.49$$

- 结论：接受H_0。不同窝别动物的出生重没有显著差异。

- 也可以将表中的每个数值都减去30：

动物编号	窝别			
	I	II	III	IV
1	4.7	3.2	-2.9	2.9
2	3.3	-4.0	-6.7	1.4
3	-3.8	-1.4	-2.2	-4.3
4	1.6	2.3	-3.3	-2.0
和	5.8	0.1	-15.1	-2.0
均值	1.450	0.025	-3.775	-0.500

$$SST = 177.52 \qquad SSA = 58.575 \qquad SSE = 118.945$$

$$F = \frac{MSA}{MSE} = \frac{SSA/df_A}{SSE/df_E} = 1.97$$

第三节　多重比较

一. 最小显著差数法（LSD检验）

*LSD*方法

1. H0: $\mu_i = \mu_j$ (第i个总体的均值等于第j个总体的均值)
2. 检验的统计量为： $\overline{x_i} - \overline{x_j}$

$$t = \frac{\overline{x_i} - \overline{x_j}}{S_{\overline{x_i} - \overline{x_j}}} \quad \Rightarrow \quad LSD = t_{\alpha/2}(n-k) \cdot S_{\overline{x_i} - \overline{x_j}}$$

$$S_{\overline{x_i} - \overline{x_j}} = \sqrt{\frac{S_i^2}{n_i} + \frac{S_j^2}{n_j}} = \sqrt{\frac{2MSE}{n_0}}$$

若$|\overline{x_i} - \overline{x_j}| \geq LSD$，拒绝$H_0$，若$|\overline{x_i} - \overline{x_j}| < LSD$，不能拒绝$H_0$

方差分析中的多重比较
（实例）

方差分析：单因素方差分析

SUMMARY

组	计数	求和	平均	方差
列 1	5	136.6	27.32	2.672
列 2	5	147.8	29.56	2.143
列 3	5	132.2	26.44	3.298
列 4	5	157.3	31.46	1.658

方差分析

差异源	SS	df	MS	F	P-value	F crit
组间	76.8455	3	25.615	10.486	0.00047	3.2389
组内	39.084	16	2.4428			

方差分析中的多重比较
（实例）

1. 根据前面的计算结果：$\bar{x}_1$=27.3; $\bar{x}_2$=29.6; $\bar{x}_3$=26.4; $\bar{x}_4$=31.5
2. 提出假设
   - H_0: $\mu_i = \mu_j$ ； H_1: $\mu_i \neq \mu_j$
3. 计算*LSD*

$$LSD = t_{0.025}(16) \cdot S_{\overline{x_i}-\overline{x_j}} = 2.12 \times \sqrt{\frac{2 \times 2.4428}{5}} = 2.096$$

方差分析中的多重比较
（实例）

$|\bar{x}_1 - \bar{x}_2| = |27.3-29.6| = 2.3 > 2.096$

颜色1与颜色2的销售量有显著差异

$|\bar{x}_1 - \bar{x}_3| = |27.3-26.4| = 0.9 < 2.096$

颜色1与颜色3的销售量没有显著差异

$|\bar{x}_1 - \bar{x}_4| = |27.3-31.5| = 4.2 > 2.096$

颜色1与颜色4的销售量有显著差异

$|\bar{x}_2 - \bar{x}_3| = |29.6-26.4| = 3.2 > 2.096$

颜色2与颜色3的销售量有显著差异

$|\bar{x}_2 - \bar{x}_4| = |29.6-31.5| = 1.9 < 2.096$

颜色2与颜色4的销售量没有显著差异

$|\bar{x}_3 - \bar{x}_4| = |26.4-31.5| = 5.1 > 2.096$

颜色3与颜色4的销售量有显著差异

组内观测次数不等的方差分析

例：园艺研究所调查了3个品种草莓的维C含量（mg/100g），测定结果如下。试分析不同品种草莓之间的维C含量是否有显著差异。

品种	维C含量									
	1	2	3	4	5	6	7	8	9	10
1	117	99	107	112	113	106				
2	81	77	79	76	85	87	74	69	72	80
3	80	82	78	84	89	73	86	88		

组内观测次数不等的方差分析

例：园艺研究所调查了3个品种草莓的维C含量（mg/100g），测定结果如下。试分析不同品种草莓之间的维C含量是否有显著差异。

品种	维C含量										合计	均值
	1	2	3	4	5	6	7	8	9	10		
1	47	29	37	42	43	36					234	39
2	11	7	9	6	15	17	4	-1	2	10	80	8
3	10	12	8	14	19	3	16	18			100	12.5

H_0: $\mu_1 = \mu_2 = \mu_3$ (三种草莓维C含量无显著差异)

H_1: μ_1、μ_2、μ_3不全相等 (有显著差异)

$$C = n\bar{\bar{x}}^2 = 7141.5$$

$$SST = \sum_{i=1}^{3}\sum_{j=1}^{n_i} x_{ij}^2 - C = 4562.5$$

$$SSA = \sum_{i=1}^{3} n_i \bar{x}_i^2 - C = 3874.5$$

$$SSE = SST - SSA = 688$$

$$F = \frac{MSA}{MSE} = \frac{SSA/df_A}{SSE/df_E}$$

$$= \frac{3874.5/2}{688/21} = 59.13$$

$$F_{0.05}(2, 21) = 3.47$$

- 结论：拒绝H_0。三个品种草莓的维C含量差异非常显著。

用LSD方法进行多重比较：

$$LSD = t_{\frac{\alpha}{2}}(n-k)\cdot S_{\overline{x_i}-\overline{x_j}} = t_{0.025}(21)\cdot S_{\overline{x_i}-\overline{x_j}}$$

$$S_{\overline{x_i}-\overline{x_j}} = \sqrt{\frac{2MSE}{n_0}}$$

$$n_0 = \frac{(\sum n_i)^2 - \sum n_i^2}{(\sum n_i)(k-1)} = 7.8 \approx 8$$

$$LSD = 2.08 \times \sqrt{\frac{2\times 32.76}{8}} = 5.95(mg/100g)$$

结　束