# 第六章　方差分析

# 适用范围

- 定类变量—定距变量
- 例如：性别与生育意愿、地区与平均寿命、民族与离婚率、职业与人际交往频次等。

- 检验总体间的均值是否有所不同，检验的方法和手段是通过方差。

- 类型：一元、多元（这里的元是指因变量）

- 方差分析的基本问题
- 单因素方差分析
- 多因素方差分析

# 方差分析的基本问题

方差分析的内容

方差分析的原理

$F$ 分布

# 什么是方差分析？

# 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

   2个或多个 ($k$ 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

# 什么是方差分析?

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$\mu_1$为无色饮料的平均销售量，$\mu_2$粉色饮料的平均销售量，$\mu_3$为橘黄色饮料的平均销售量，$\mu_4$为绿色饮料的平均销售量，也就是检验下面的假设
   - $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
   - $H_1$: $\mu_1$ , $\mu_2$ , $\mu_3$ , $\mu_4$ 不全相等
3. 检验上述假设所采用的方法就是方差分析

# 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. 因素或因子
   - 所要检验的对象称为因子
   - 要分析饮料的颜色对销售量是否有影响，**颜色**是要检验的因素或因子
2. 水平
   - 因素的具体表现称为水平
   - $A_1$、$A_2$、$A_3$、 $A_4$四种颜色就是因素的水平
3. 观察值
   - 在每个因素水平下得到的样本值
   - 每种颜色饮料的销售量就是观察值

# 方差分析的基本思想和原理

1. 试验
   - 这里只涉及一个因素，因此称为单因素四水平的试验
2. 总体
   - 因素的每一个水平可以看作是一个总体
   - 比如$A_1$、$A_2$、$A_3$、 $A_4$四种颜色可以看作是四个总体
3. 样本数据
   - 上面的数据可以看作是从这四个总体中抽取的样本数据

# 方差分析的基本思想和原理

1. 比较两类误差，以检验均值是否相等
2. 比较的基础是方差比
3. 如果系统(处理)误差显著地不同于随机误差，则均值就是不相等的；反之，均值就是相等的
4. 误差是由各部分的误差占总误差的比例来测度的

# 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

# 方差分析中的基本假定

1. 在上述假定条件下，判断颜色对销售量是否有显著影响，实际上也就是检验具有同方差的四个正态总体的均值是否相等的问题
2. 如果四个总体的均值相等，可以期望四个样本的均值也会很接近
   - 四个样本的均值越接近，我们推断四个总体均值相等的证据也就越充分
   - 样本均值越不同，我们推断总体均值不同的证据就越充分

# 方差分析中基本假定

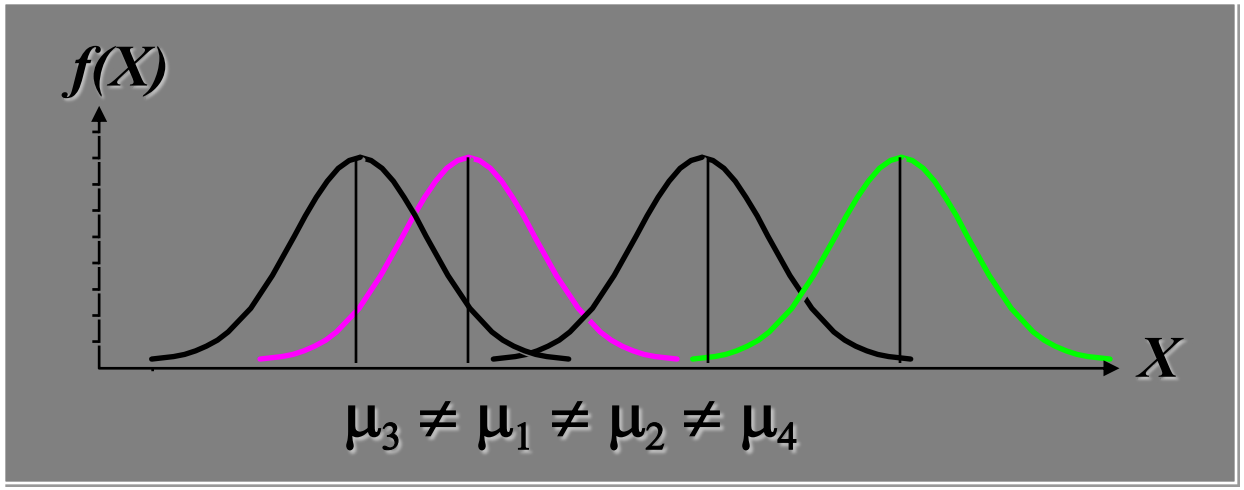
- 如果原假设成立，即$H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
  - 四种颜色饮料销售的均值都相等
  - 没有系统误差

这意味着每个样本都来自均值为 $\mu$、差为$\sigma^2$的同一正态总体

# 方差分析中基本假定

- 如果备择假设成立，即$H_1$: $\mu_i$ ($i$=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



# 单因素方差分析

单因素方差分析的步骤

方差分析中的多重比较

单因素方差分析中的其他问题

# 单因素方差分析的数据结构

| 观察值（*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* | $x_{n1}$ | $x_{n2}$ | … | $x_{nk}$ |

# 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）
   - $H_1$: $\mu_1$ ， $\mu_2$ ， … ， $\mu_k$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
   - $H_0$: $\mu_1$ ， $\mu_2$ ， $\mu_3$， $\mu_4$不全相等
     - 颜色对销售量有影响

# 构造检验的统计量

1. 为检验$H_0$是否成立，需确定检验的统计量
2. 构造统计量需要计算
   - 水平的均值
   - 全部观察值的总均值
   - 离差平方和
   - 均方($MS$)

# 构造检验的统计量
# (计算水平的均值 )

1. 假定从第$i$个总体中抽取一个容量为$n_i$的简单随机样本，第$i$个总体的样本均值为该样本的全部观察值总和除以观察值的个数
2. 计算公式为

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

式中：$n_i$为第 $i$ 个总体的样本观察值个数

$x_{ij}$ 为第 $i$ 个总体的第 $j$ 个观察值

# 构造检验的统计量
## (计算全部观察值的总均值 )

1. 全部观察值的总和除以观察值的总个数
2. 计算公式为

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

式中：$n = n_1 + n_2 + \cdots + n_k$

# 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 |

# 构造检验的统计量
## (计算总离差平方和 *SST*)

1. 全部观察值 $x_{ij}$ 与总平均值 $\bar{\bar{x}}$ 的离差平方和
2. 反映全部观察值的离散状况
3. 其计算公式为

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

- 前例的计算结果：

$SST = (26.5\text{-}28.695)^2+(28.7\text{-}28.695)^2+\ldots+(32.8\text{-}28.695)^2$

$=115.9295$

# 构造检验的统计量
# (计算误差项平方和 *SSE*)

1. 每个水平或组的各样本数据与其组平均值的离差平方和
2. 反映每个样本各观察值的离散状况，又称组内离差平方和
3. 该平方和反映的是随机误差的大小
4. 计算公式为

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

- 前例的计算结果：$SSE = 39.084$

# 构造检验的统计量
# (计算水平项平方和 *SSA*)

1. 各组平均值 $\bar{x}_i$ $(i = 1,2,\cdots,k)$与总平均值 $\bar{\bar{x}}$的离差平方和
2. 反映各总体的样本均值之间的差异程度，又称组间平方和
3. 该平方和既包括随机误差，也包括系统误差
4. 计算公式为

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

- 前例的计算结果：$SSA = 76.8455$

# 构造检验的统计量
# (三个平方和的关系)

- 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

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

$$SST = SSE + SSA$$

# 构造检验的统计量
## (三个平方和的作用)

1. ***SST***反映了全部数据总的误差程度；***SSE***反映了随机误差的大小；***SSA***反映了随机误差和系统误差的大小
2. 如果原假设成立，即$H_1 = H_2 = \ldots = H_k$为真，则表明没有系统误差，组间平方和*SSA*除以自由度后的**均方**与组内平方和*SSE*和除以自由度后的**均方**差异就不会太大；如果**组间均方**显著地大于**组内均方**，说明各水平(总体)之间的差异不仅有随机误差，还有系统误差
3. 判断因素的水平是否对其观察值有影响，实际上就是比较**组间方差**与**组内方差**之间差异的大小
4. 为检验这种差异，需要构造一个用于检验的统计量

# 构造检验的统计量
# (计算均方 *MS*)

1. 各离差平方和的大小与观察值的多少有关，为了消除观察值多少对离差平方和大小的影响，需要将其平均，这就是均方，也称为方差
2. 计算方法是用离差平方和除以相应的自由度
3. 三个平方和的自由度分别是
   - ***SST*** 的自由度为$n$-1，其中$n$为全部观察值的个数
   - ***SSA***的自由度为$k$-1，其中$k$为因素水平(总体)的个数
   - ***SSE*** 的自由度为$n$-$k$

# 构造检验的统计量
## (计算均方 *MS*)

1. *SSA*的均方也称**组间方差**，记为*MSA*，计算公式为

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

前例的计算结果：$MSA = \frac{76.8455}{4-1} = 25.6152$

2. *SSE*的均方也称**组内方差**，记为*MSE*，计算公式为

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

前例的计算结果：$MSE = \frac{39.084}{20-4} = 2.4428$

# 构造检验的统计量
## (计算检验的统计量 $F$)

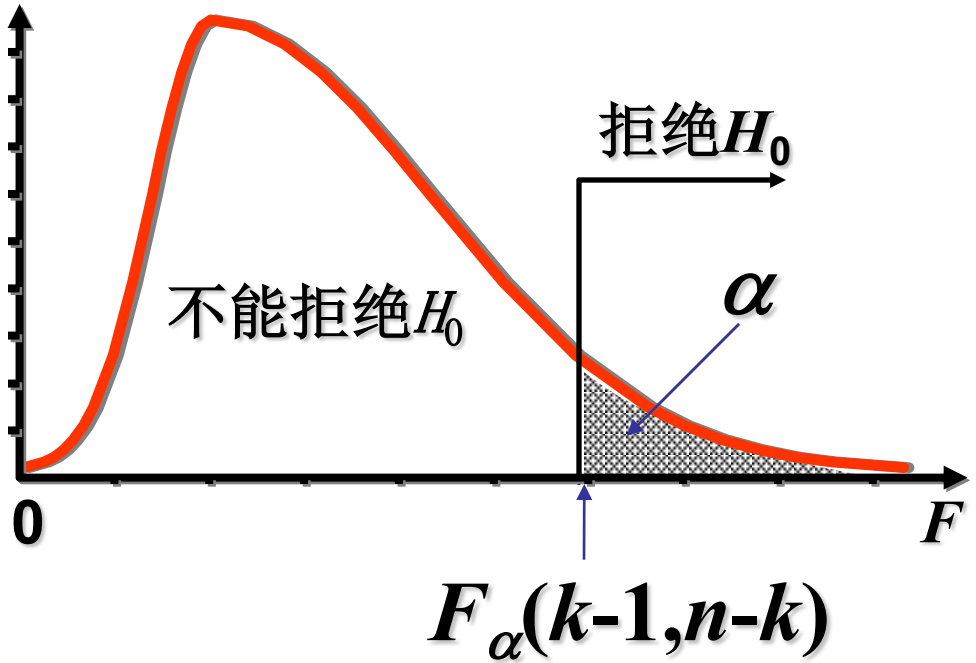
1. 将$MSA$和$MSE$进行对比，即得到所需要的检验统计量$F$
2. 当$H_0$为真时，二者的比值服从分子自由度为$k$-1、分母自由度为 $n$-$k$ 的 $F$ 分布，即

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

前例的计算结果：$F = \frac{25.6152}{2.4428} = 10.486$

# 构造检验的统计量
## (*F*分布与拒绝域)

如果均值相等，
$F=MSA/MSE \to 1$




*F* 分布

# 统计决策

- 将统计量的值$F$与给定的显著性水平$\alpha$的临界值$F_\alpha$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$df_1=k-1$、第二自由度$df_2=n-k$ 相应的临界值 $F_\alpha$
  - 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$F\leq F_\alpha$ ，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

# 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 | | | |
| | | | | | | |
| 总计 | 115.93 | 19 | | | | |

【例】为了对几个行业的服务质量进行评价，消费者协会在零售业、旅游业、航空公司、家电制造业分别抽取了不同的样本，其中零售业抽取7家，旅游业抽取了6家，航空公司抽取5家、家电制造业抽取了5家，然后记录了一年中消费者对总共23家服务企业投诉的次数，结果如下表。试分析这四个行业的服务质量是否有显著差异？($\alpha=0.05$)

**消费者对四个行业的投诉次数**

| 观察值 ($j$) | 行业($A$) | | | |
|---|---|---|---|---|
| | 零售业 | 旅游业 | 航空公司 | 家电制造业 |
| 1 | 57 | 62 | 51 | 70 |
| 2 | 55 | 49 | 49 | 68 |
| 3 | 46 | 60 | 48 | 63 |
| 4 | 45 | 54 | 55 | 69 |
| 5 | 54 | 56 | 47 | 60 |
| 6 | 53 | 55 | | |
| 7 | 47 | | | |

# （计算结果）

- 解：设四个行业被投诉次数的均值分别为，$\mu_1$、$\mu_2$ 、$\mu_3$、$\mu_4$，则需要检验如下假设
  - $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (四个行业的服务质量无显著差异)
  - $H_1$: $\mu_1$，$\mu_2$，$\mu_3$，$\mu_4$不全相等 (有显著差异)
  - Excel输出的结果如下

| 差异源 | SS | 自由度 | MS | F | P-值 | 临界值 |
|---|---|---|---|---|---|---|
| 组间 | 845.2174 | 3 | 281.7391 | 14.78741 | 3.31E-05 | 3.127354 |
| 组内 | 362 | 19 | 19.05263 | | | |
| 总和 | 1207.217 | 22 | | | | |

- 结论：拒绝$H_0$。四个行业的服务质量有显著差异

# 一元方差分析

- 职业与家庭子女数
- 地区与平均家庭人口
- 科研单位与拥有高级技术人才数量
- ……

# 方差分析中的多重比较

# 方差分析中的多重比较（作用）

1. 多重比较是通过对总体均值之间的配对比较来进一步检验到底哪些均值之间存在差异
2. 多重比较方法有多种，这里介绍Fisher提出的**最小显著差异**方法，简写为***LSD***，该方法可用于判断到底哪些均值之间有差异
3. *LSD*方法是对检验两个总体均值是否相等的*t*检验方法的总体方差估计加以修正(用*MSE*来代替)而得到的

# 方差分析中的多重比较
## （步骤）

1. 提出假设
   - $H_0$: $\mu_i = \mu_j$ (第$i$个总体的均值等于第$j$个总体的均值)
   - $H_1$: $\mu_i \neq \mu_j$ (第$i$个总体的均值不等于第$j$个总体的均值)
2. 检验的统计量为

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim t(n-k)$$

3. 若$|t| \geq t_{\alpha/2}$，拒绝$H_0$；若$|t| < t_{\alpha/2}$，不能拒绝$H_0$

# 方差分析中的多重比较
## （基于统计量 $\bar{x}_i - \bar{x}_j$的*LSD*方法）

1. 通过判断样本均值之差的大小来检验 $H_0$
2. 检验的统计量为： $\bar{x}_i - \bar{x}_j$
3. 检验的步骤为
   - 提出假设
     - $H_0$: $\mu_i = \mu_j$ (第*i*个总体的均值等于第*j*个总体的均值)
     - $H_1$: $\mu_i \neq \mu_j$ (第*i*个总体的均值不等于第*j*个总体的均值)
   - 计算*LSD*

$$LSD = t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

   - 若$|\bar{x}_i - \bar{x}_j| \geq LSD$，拒绝$H_0$，若$|\bar{x}_i - \bar{x}_j| < LSD$，不能拒绝$H_0$

# 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; $\bar{x}_2$=29.5; $\bar{x}_3$=26.4; $\bar{x}_4$=31.4
2. 提出假设
   - $H_0$: $\mu_i = \mu_j$ ； $H_1$: $\mu_i \neq \mu_j$
3. 计算*LSD*

$$LSD = 2.12\sqrt{2.4428\left(\frac{1}{5}+\frac{1}{5}\right)} = 2.096$$

# 方差分析中的多重比较
## （实例）

$|\bar{x}_1 - \bar{x}_2| = |27.3-29.5| = 2.2 > 2.096$

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

$|\bar{x}_1 - \bar{x}_4| = |27.3-31.4| = 4.1 > 2.096$

颜色1与颜色4的销售量有显著差异

$|\bar{x}_2 - \bar{x}_3| = |29.5-26.4| = 3.1 > 2.096$

颜色2与颜色3的销售量有显著差异

$|\bar{x}_2 - \bar{x}_4| = |29.5-31.4| = 1.9 < 2.096$

颜色2与颜色4的销售量没有显著差异

$|\bar{x}_3 - \bar{x}_4| = |26.4-31.4| = 5 > 2.096$

颜色3与颜色4的销售量有显著差异

# 双因素方差分析

双因素方差分析的基本问题

双因素方差分析的数据结构

双因素方差分析的步骤

一个应用实例

# 双因素方差分析的基本问题

# 数学模型

- 线性可加模型（独立模型）

  Yij=Y平均+Ai的效果+Bj 的效果+eij

  观测1次

- 交互作用模型：

  Yij=Y平均+Ai的效果+Bj 的效果+（AB)ij交互作用+eij

  每种情况至少观测2次

# 例子

- 教学方法与教师性格对教学效果的影响

- 注入式A1、启发式A2
- 内向型B1 、外向型B2

# 双因素方差分析
## （概念要点）

1. 分析两个因素(因素$A$和因素$B$)对试验结果的影响
2. 分别对两个因素进行检验，分析是一个因素在起作用，还是两个因素都起作用，还是两个因素都不起作用
3. 如果$A$和$B$对试验结果的影响是相互独立的，分别判断因素$A$和因素$B$对试验指标的影响，这时的双因素方差分析称为无交互作用的双因素方差分析
4. 如果除了$A$和$B$对试验结果的单独影响外，因素$A$和因素$B$的搭配还会对销售量产生一种新的影响，这时的双因素方差分析称为有交互作用的双因素方差分析
5. 对于无交互作用的双因素方差分析，其结果与对每个因素分别进行单因素方差分析的结果相同

# 双因素方差分析的基本假定

1. 每个总体都服从正态分布
   - 对于因素的每一个水平，其观察值是来自正态分布总体的简单随机样本
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
3. 观察值是独立的

# 无交互双因素方差分析的数据结构

| 因素$A$ ($i$) | 因素($B$) $j$ | | | | 平均值 |
|---|---|---|---|---|---|
| | $B_1$ | $B_2$ | … | $B_r$ | $\bar{x}_{i.}$ |
| $A_1$ | $x_{11}$ | $x_{12}$ | … | $x_{1k}$ | $\bar{x}_{1.}$ |
| $A_2$ | $x_{21}$ | $x_{22}$ | … | $x_{2k}$ | $\bar{x}_{2.}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $A_k$ | $x_{r1}$ | $x_{r2}$ | … | $x_{rk}$ | $\bar{x}_{k.}$ |
| 平均值 $\bar{x}_{.j}$ | $\bar{x}_{.1}$ | $\bar{x}_{.2}$ | … | $\bar{x}_{.r}$ | $\bar{\bar{x}}$ |

$$x_{ij}\ (i=1,2,\cdots,k;\ j=1,2,\cdots,r)$$

- $\bar{x}_{i\cdot}$是因素$A$的第$i$个水平下各观察值的平均值

$$\bar{x}_{i\cdot}=\frac{\sum_{j=1}^{r}x_{ij}}{r}\qquad(i=1,2,\cdots,k)$$

- $\bar{x}_{\cdot j}$是因素$B$的第$j$个水平下的各观察值的均值

$$\bar{x}_{\cdot j}=\frac{\sum_{i=1}^{k}x_{ij}}{k}\qquad(j=1,2,\cdots,r)$$

- $\bar{\bar{x}}$是全部 $kr$ 个样本数据的总平均值

$$\bar{\bar{x}}=\frac{\sum_{i=1}^{k}\sum_{j=1}^{r}x_{ij}}{kr}$$

# 双因素方差分析的步骤

# 提出假设

1. 对因素$A$提出的假设为
   - $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_k$ ($\mu_i$为第$i$个水平的均值)
   - $H_1$: $\mu_i$ ($i$ =1,2, … , $k$) 不全相等
2. 对因素$B$提出的假设为
   - $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_r$ ($\mu_j$为第$j$个水平的均值)
   - $H_1$: $\mu_j$ ($j$ =1,2,…,$r$) 不全相等

# 构造检验的统计量

1. 为检验$H_0$是否成立，需确定检验的统计量
2. 构造统计量需要计算
   - 总离差平方和
   - 水平项平方和
   - 误差项平方和
   - 均方

# 构造检验的统计量
## (计算总离差平方和 *SST*)

1. 全部观察值 $x_{ij}$ $(i=1,2,\cdots,k;\ j=1,2,\cdots,r)$与总平均值 $\bar{\bar{x}}$ 的离差平方和
2. 反映全部观察值的离散状况
3. 计算公式为

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

# 构造检验的统计量
## (计算*SSA*、*SSB*和*SSE*)

1. 因素*A*的离差平方和***SSA***

$$SSA=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{i\cdot}-\bar{\bar{x}}\right)^2$$

2. 因素*B*的离差平方和***SSB***

$$SSB=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{\cdot j}-\bar{\bar{x}}\right)^2$$

3. 误差项平方和***SSE***

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{\bar{x}}\right)^2$$

# 构造检验的统计量
## (各平方和的关系)

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

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

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

$$\boldsymbol{SST = SSA + SSB + SSE}$$

# 构造检验的统计量
## (计算均方 *MS*)

1. 各离差平方和的大小与观察值的多少有关，为消除观察值多少对离差平方和大小的影响，需要将其平均，这就是均方，也称为方差
2. 计算方法是用离差平方和除以相应的自由度
3. 三个平方和的自由度分别是
   - 总离差平方和*SST*的自由度为 $kr-1$
   - 因素*A*的离差平方和*SSA*的自由度为 $k-1$
   - 因素*B*的离差平方和*SSB*的自由度为 $r-1$
   - 随机误差平方和*SSE*的自由度为 $(k-1)\times(r-1)$

# 构造检验的统计量
## (计算均方 *MS*)

1. 因素*A*的均方，记为*MSA*，计算公式为

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

2. 因素*B*的均方，记为*MSB*，计算公式为

$$MSB = \frac{SSB}{r-1}$$

3. 随机误差项的均方，记为*MSE*，计算公式为

$$MSE = \frac{SSE}{(k-1)(r-1)}$$

# 构造检验的统计量
## (计算检验的统计量 $F$)

1. 为检验因素$A$的影响是否显著，采用下面的统计量

$$F_A = \frac{MSA}{MSE} \sim F\left(k-1,(k-1)(r-1)\right)$$

2. 为检验因素$B$的影响是否显著，采用下面的统计量

$$F_B = \frac{MSB}{MSE} \sim F\left(r-1,(k-1)(r-1)\right)$$

# 统计决策

- 将统计量的值$F$与给定的显著性水平$\alpha$的临界值$F_\alpha$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$\alpha$在$F$分布表中查找相应的临界值 $F_\alpha$
  - 若$F_A \geq F_\alpha$ ，则拒绝原假设$H_0$ ，表明均值之间的差异是显著的，即所检验的因素($A$)对观察值有显著影响
  - 若$F_B \geq F_\alpha$ ，则拒绝原假设$H_0$ ，表明均值之间有显著差异，即所检验的因素($B$)对观察值有显著影响

# 双因素方差分析表

| 方差来源 | 平方和<br>*SS* | 自由度<br>*df* | 均方<br>*MS* | *F*值 |
|---|---|---|---|---|
| 因素*A* | *SSA* | *k*-1 | *MSA* | $F_A$ |
| 因素*B* | *SSB* | *r*-1 | *MSB* | $F_B$ |
| 误差 | *SSE* | (*k*-1)×(r-1) | *MSE* | |
| 总和 | *SST* | *kr*-1 | | |

# 例子

有四个品牌的彩电在五个地区销售，为分析彩电的品牌(因素$A$)和销售地区(因素$B$)对销售量是否有影响，对每个品牌在各地区的销售量取得以下数据，见下表。试分析品牌和销售地区对彩电的销售量是否有显著影响？

**不同品牌的彩电在各地区的销售量数据**

| 品牌 (因素$A$) | 销售地区( 因素$B$ ) | | | | |
|---|---|---|---|---|---|
| | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ |
| $A_1$ | 365 | 350 | 343 | 340 | 323 |
| $A_2$ | 345 | 368 | 363 | 330 | 333 |
| $A_3$ | 358 | 323 | 353 | 343 | 308 |
| $A_4$ | 288 | 280 | 298 | 260 | 298 |

# 双因素方差分析
## （提出假设）

1. 对因素$A$提出的假设为
   - $H_0$: $\mu_1=\mu_2=\mu_3=\mu_4$
     (品牌对销售量没有影响)
   - $H_1$: $\mu_i$ ($i$ =1,2, … , 4) 不全相等
     (品牌对销售量有影响)
2. 对因素$B$提出的假设为
   - $H_0$: $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$
     (地区对销售量没有影响)
   - $H_1$: $\mu_j$ ($j$ =1,2,…,5) 不全相等
     (地区对销售量有影响)

| 差异源 | SS | df | MS | F | P-value | F crit |
|---|---|---|---|---|---|---|
| 行(因素 A) | 13004.55 | 3 | 4334.85 | 18.10777 | 9.46E-05 | 3.4903 |
| 列(因素 B) | 2011.7 | 4 | 502.925 | 2.100846 | 0.143665 | 3.2592 |
| 误差 | 2872.7 | 12 | 239.3917 | | | |
| 总和 | 17888.95 | 19 | | | | |

结论：

- $F_A = 18.10777 > F_\alpha = 3.4903$，拒绝原假设$H_0$，说明彩电的品牌对销售量有显著影响
- $F_B = 2.100846 < F_\alpha = 3.2592$，接受原假设$H_0$，说明销售地区对彩电的销售量没有显著影响

# 带有交互作用的双因素方差分析

- 数据需要观测2次或多次
- 原理

# （2）有交互作用的双因素方差分析

为了研究两个因素是否独立，有无交互作用，我们需要在各个因素水平组合下，进行重复试验。

**表 7.13：双因素重复试验数据**

| | $B_1$ | $B_2$ | $\cdots$ | $B_s$ |
|---|---|---|---|---|
| $A_1$ | $x_{111}x_{112}\cdots x_{11t}$ | $x_{121}x_{122}\cdots x_{12t}$ | $\cdots$ | $x_{1s1}x_{1s2}\cdots x_{1st}$ |
| $A_2$ | $x_{211}x_{212}\cdots x_{21t}$ | $x_{221}x_{222}\cdots x_{22t}$ | $\cdots$ | $x_{2s1}x_{2s2}\cdots x_{2st}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| $A_r$ | $x_{r11}x_{r12}\cdots x_{r1t}$ | $x_{r21}x_{r22}\cdots x_{r2t}$ | $\cdots$ | $x_{rs1}x_{rs2}\cdots x_{rst}$ |

**1. 数学模型**

设有两个因素 $A$ 和 $B$, 因素 $A$ 有 $r$ 个水平 $A_1, A_2, \cdots, A_r$; 因素 $B$ 有 $s$ 个水平 $B_1, B_2, \cdots, B_s$, 每种水平组合 $(A_i, B_j)$ 下重复试验 $t$ 次. 记第 $k$ 次的观测值为 $x_{ijk}$, 将观测数据列表, 如表 7.13 所示.

假定

$$x_{ijk} \sim N(\mu_{ij}, \sigma^2), \quad i = 1, 2, \ldots, r; \ j = 1, 2, \ldots, s; \ k = 1, 2, \ldots, t,$$

各 $x_{ijk}$ 相互独立. 所以, 数据可以分解为

$$\begin{cases} x_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} \sim N(0, \sigma^2), \text{且各 } \varepsilon_{ijk} \text{ 相互独立}, \\ i = 1, 2, \cdots, r, \quad j = 1, 2, \cdots, s, \quad k = 1, 2, \ldots, t, \end{cases} \tag{7.18}$$

其中 $\alpha_i$ 为因素 $A$ 的第 $i$ 个水平的效应, $\beta_j$ 为因素 $B$ 的第 $j$ 个水平的效应. $\delta_{ij}$ 表示 $A_i$ 和 $B_j$ 的交互效应, 因此有

$$\mu = \frac{1}{rs} \sum_{i=1}^{r} \sum_{j=1}^{s} \mu_{ij}, \quad \sum_{i=1}^{r} \alpha_i = 0, \quad \sum_{j=1}^{s} \beta_j = 0, \quad \sum_{i=1}^{r} \delta_{ij} = \sum_{j=1}^{s} \delta_{ij} = 0.$$

**2. 方差分析**

此时判断因素 $A, B$ 及交互效应的影响是否显著等价于检验下列假设

$$
\begin{aligned}
H_{01}: &\quad \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0, \\
H_{02}: &\quad \beta_1 = \beta_2 = \cdots = \beta_r = 0, \\
H_{03}: &\quad \delta_{ij} = 0, \quad i = 1, 2, \cdots, r, \quad j = 1, 2, \cdots s.
\end{aligned}
$$

在这种情况下，方差分析法与前两节的方法类似，有下列计算公式：

$$
S_T = S_E + S_A + S_B + S_{A\times B},
$$

其中

$$
\begin{aligned}
S_T &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk} - \bar{x})^2, \quad \bar{x} = \frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t} x_{ijk}, \\
S_E &= \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk} - \bar{x}_{ij.})^2, \\
\bar{x}_{ij.} &= \frac{1}{t}\sum_{k=1}^{t} x_{ijk}, \quad i = 1, 2, \cdots, r, \quad j = 1, 2, \ldots, s,
\end{aligned}
$$

$$
\begin{aligned}
S_A &= st\sum_{i=1}^{r}(\bar{x}_{i\cdot\cdot}-\bar{x})^2, \quad \bar{x}_{i\cdot\cdot} = \frac{1}{st}\sum_{j=1}^{s}\sum_{k=1}^{t} x_{ijk}, \quad i=1,2,\cdots,r,\\
S_B &= rt\sum_{j=1}^{s}(\bar{x}_{\cdot j\cdot}-\bar{x})^2, \quad \bar{x}_{\cdot j\cdot} = \frac{1}{rt}\sum_{i=1}^{r}\sum_{k=1}^{t} x_{ijk}, \quad j=1,2,\cdots,s,\\
S_{A\times B} &= t\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{ij\cdot}-\bar{x}_{i\cdot\cdot}-\bar{x}_{\cdot j\cdot}+\bar{x})^2,
\end{aligned}
$$

其中 $S_T$ 为总离差平方和， $S_E$ 为误差平方和， $S_A$ 为因素 $A$ 的平方和， $S_B$ 为因素 $B$ 的平方和， $S_{A\times B}$ 为交互效应平方和. 可以证明：当 $H_{01}$ 成立时，

$$
F_A = \frac{S_A/(r-1)}{S_E/[rs(t-1)]} \sim F(r-1, rs(t-1)).
$$

当 $H_{02}$ 成立时，

$$
F_B = \frac{S_B/(s-1)}{S_E/[rs(t-1)]} \sim F(s-1, rs(t-1)).
$$

当 $H_{03}$ 成立时，

$$
F_{A\times B} = \frac{S_{A\times B}/[(r-1)(s-1)]}{S_E/[rs(t-1)]} \sim F((r-1)(s-1), rs(t-1)).
$$

**表 7.14: 有交互效应的双因素方差分析表**

| 方差来源 | 自由度 | 平方和 | 均方 | F 比 | P- 值 |
|---|---|---|---|---|---|
| 因素 A | $r-1$ | $S_A$ | $MS_A = \frac{S_A}{r-1}$ | $F_A = \frac{MS_A}{MS_E}$ | $p_A$ |
| 因素 B | $s-1$ | $S_B$ | $MS_B = \frac{S_B}{s-1}$ | $F_B = \frac{MS_B}{MS_E}$ | $p_B$ |
| 交互效应 $A \times B$ | $(r-1)(s-1)$ | $S_{A\times B}$ | $MS_{A\times B} = \frac{S_{A\times B}}{(r-1)(s-1)}$ | $F_{A\times B} = \frac{MS_{A\times B}}{MS_E}$ | $p_{A\times B}$ |
| 误 差 | $rs(t-1)$ | $S_E$ | $MS_E = \frac{S_E}{rs(t-1)}$ | | |
| 总 和 | $rst-1$ | $S_T$ | | | |

# 多因素方差分析

- 模型：
  - 饱和模型、不饱和模型
  - 主效应、交互效应（各阶）
- 数据：要重复观测
- 原理（略）

# 多元方差分析

- 案例来自随机抽样
- 各因变量为正态分布且方差相等
- 各因变量之间为多元正态分布

- 因变量之间具有足够的相关性
- 单因素、多因素

# 方差分析条件的检验

- 正态条件检验
- 方差齐性检验