第10章　方差分析

PowerPoint

第10章　方差分析

10.1　方差分析引论
10.2　单因素方差分析
10.3　双因素方差分析

学习目标

1. 解释方差分析的概念
2. 解释方差分析的基本思想和原理
3. 掌握单因素方差分析的方法及应用
4. 理解多重比较的意义
5. 掌握双因素方差分析的方法及应用

10.1 方差分析引论

10.1.1 方差分析及其有关术语
10.1.2 方差分析的基本思想和原理
10.1.3 方差分析的基本假定
10.1.4 问题的一般提法

作者：贾俊平，中国人民大学统计学院

方差分析及其有关术语

作者：贾俊平，中国人民大学统计学院

什么是方差分析(ANOVA)? (analysis of variance)

1. 检验多个总体均值是否相等
   - 通过分析数据的误差判断各总体均值是否相等
2. 研究分类型自变量对数值型因变量的影响
   - 一个或多个分类型自变量
     - 两个或多个 (k 个) 处理水平或分类
   - 一个数值型因变量
3. 有单因素方差分析和双因素方差分析
   - 单因素方差分析：涉及一个分类的自变量
   - 双因素方差分析：涉及两个分类的自变量

什么是方差分析?
(例题分析)

【例】为了对几个行业的服务质量进行评价，消费者协会在4个行业分别抽取了不同的企业作为样本。最近一年中消费者对总共23家企业投诉的次数如下表

消费者对四个行业的投诉次数				
	行业			
观测值	零售业	旅游业	航空公司	家电制造业
1	57	68	31	44
2	66	39	49	51
3	49	29	21	65
4	40	45	34	77
5	34	56	40	58
6	53	51		
7	44			




什么是方差分析?
(例题分析)

1. 分析4个行业之间的服务质量是否有显著差异，也就是要判断“行业”对“投诉次数”是否有显著影响
2. 作出这种判断最终被归结为检验这四个行业被投诉次数的均值是否相等
3. 若它们的均值相等，则意味着“行业”对投诉次数是没有影响的，即它们之间的服务质量没有显著差异；若均值不全相等，则意味着“行业”对投诉次数是有影响的，它们之间的服务质量有显著差异

方差分析中的有关术语

1. 因素或因子(factor)
   - 所要检验的对象
     - 分析行业对投诉次数的影响，行业是要检验的因子
2. 水平或处理(treatment)
   - 因子的不同表现
     - 零售业、旅游业、航空公司、家电制造业
3. 观察值
   - 在每个因素水平下得到的样本数据
     - 每个行业被投诉的次数

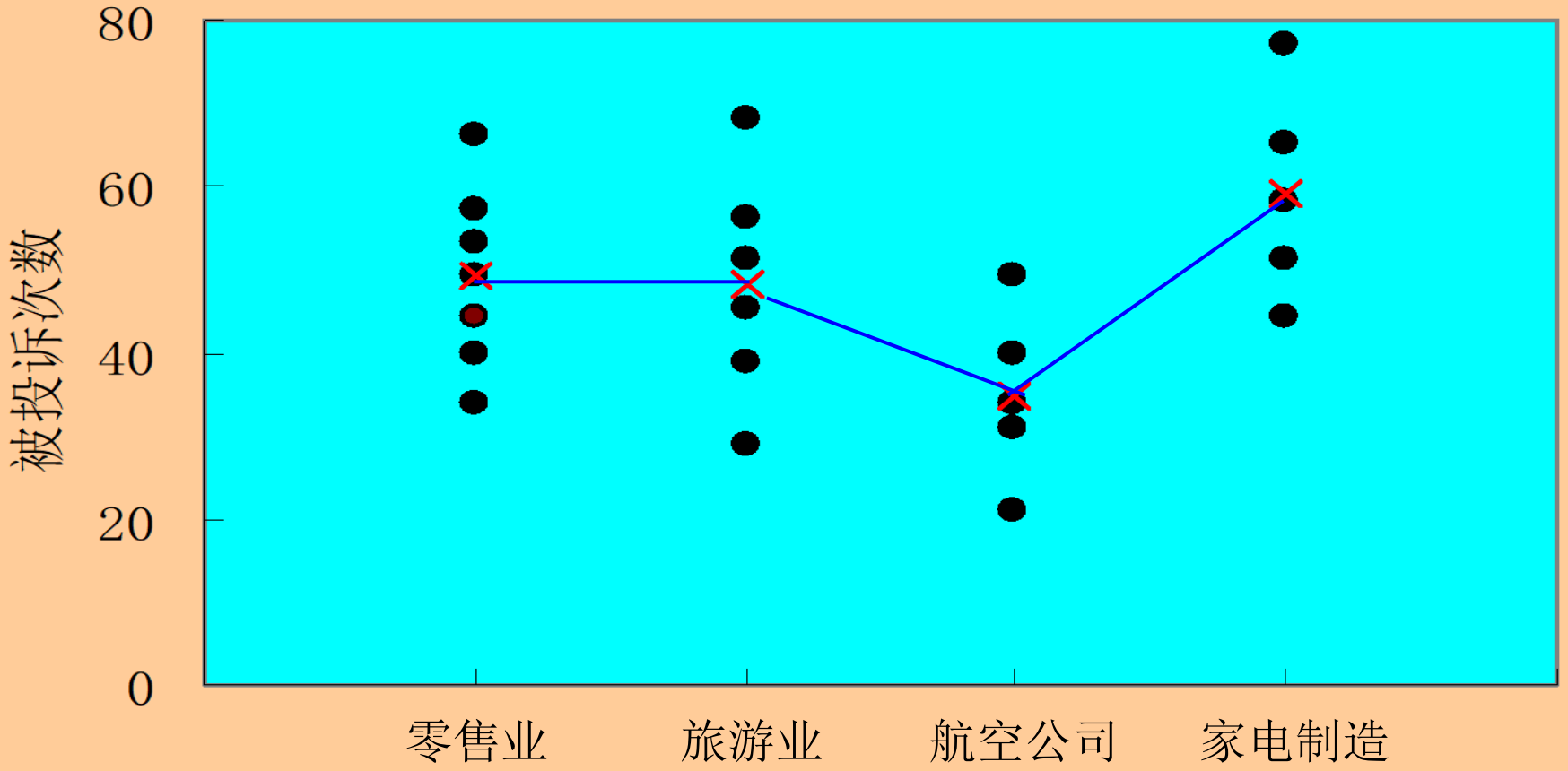
方差分析中的有关术语

1. 试验
   - 这里只涉及一个因素，因此称为单因素4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
     - 零售业、旅游业、航空公司、家电制造业是4个总体
3. 样本数据
   - 被投诉次数可以看作是从这4个总体中抽取的样本数据

方差分析的基本思想和原理

作者：贾俊平，中国人民大学统计学院

方差分析的基本思想和原理
(图形分析—散点图)



不同行业被投诉次数的散点图

方差分析的基本思想和原理 (图形分析)

1. 从散点图上可以看出
   - 不同行业被投诉的次数有明显差异
   - 同一个行业，不同企业被投诉的次数也明显不同
     - 家电制造被投诉的次数较高，航空公司被投诉的次数较低
2. 行业与被投诉次数之间有一定的关系
   - 如果行业与被投诉次数之间没有关系，那么它们被投诉的次数应该差不多相同，在散点图上所呈现的模式也就应该很接近

方差分析的基本思想和原理

1. 散点图观察不能提供充分的证据证明不同行业被投诉的次数之间有显著差异
   - 这种差异可能是由于抽样的随机性所造成的
2. 需要有更准确的方法来检验这种差异是否显著，也就是进行方差分析
   - 所以叫方差分析，因为虽然我们感兴趣的是均值，但在判断均值之间是否有差异时则需要借助于方差
   - 这个名字也表示：它是通过对数据误差来源的分析判断不同总体的均值是否相等。因此，进行方差分析时，需要考察数据误差的来源

方差分析的基本思想和原理 (两类误差)

1. 随机误差
   - 因素的同一水平(总体)下，样本各观察值之间的差异
     - 比如，同一行业下不同企业被投诉次数之间的差异
   - 这种差异可以看成是随机因素的影响，称为*随机误差*
2. 系统误差
   - 因素的不同水平(不同总体)之间观察值的差异
     - 比如，不同行业之间的被投诉次数之间的差异
   - 这种差异*可能*是由于抽样的随机性所造成的，*也可能*是由于行业本身所造成的，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 (误差平方和—SS)

1. 数据的误差用平方和(sum of squares)表示
2. 组内平方和(within groups)
   - 因素的同一水平下数据误差的平方和
     - 比如，零售业被投诉次数的误差平方和
   - 只包含*随机误差*
3. 组间平方和(between groups)
   - 因素的不同水平之间数据误差的平方和
     - 比如，4个行业被投诉次数之间的误差平方和
   - 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 (均方—MS)

1. 平方和除以相应的自由度
2. 若原假设成立，组间均方与组内均方的数值就应该很接近，它们的比值就会接近1
3. 若原假设不成立，组间均方会大于组内均方，它们之间的比值就会大于1
4. 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，即自变量对因变量有影响
   - 判断行业对投诉次数是否有显著影响，也就是检验被投诉次数的差异主要是由于什么原因所引起的。如果这种差异主要是系统误差，说明不同行业对投诉次数有显著影响

方差分析的基本假定

作者：贾俊平，中国人民大学统计学院

方差分析的基本假定

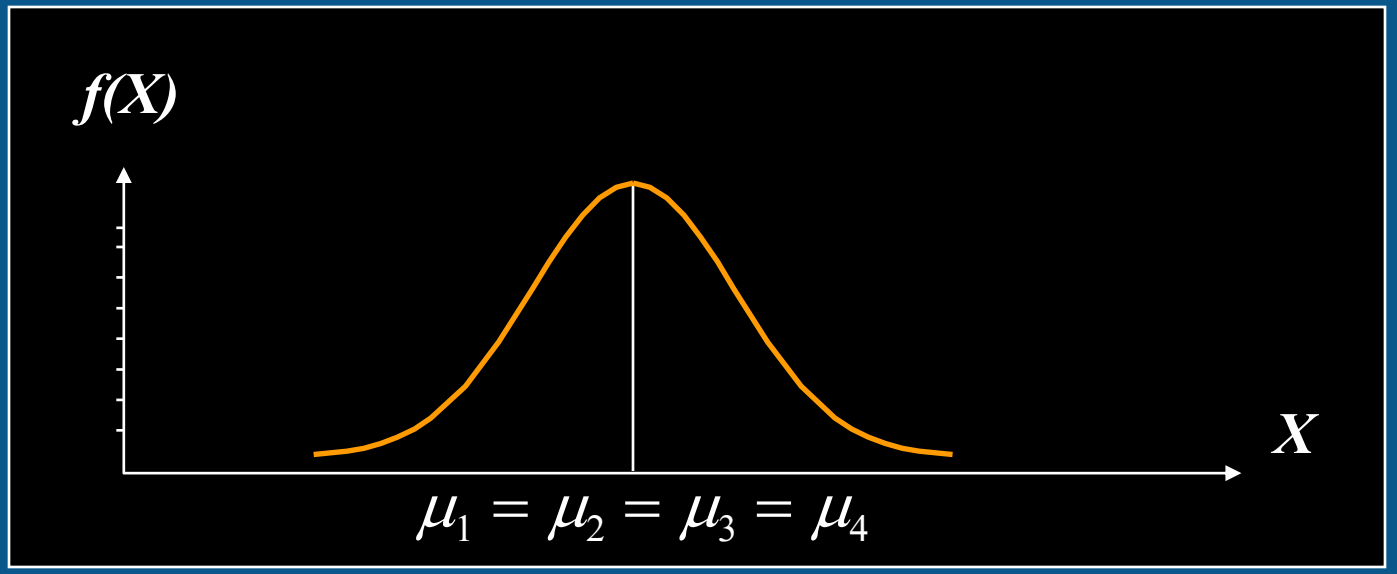
1. 每个总体都应服从正态分布
   - 对于因素的每一个水平，其观察值是来自服从正态分布总体的简单随机样本
   - 比如，每个行业被投诉的次数必须服从正态分布
2. 各个总体的方差必须相同
   - 各组观察数据是从具有相同方差的总体中抽取的
   - 比如，4个行业被投诉次数的方差都相等
3. 观察值是独立的
   - 比如，每个行业被投诉的次数与其他行业被投诉的次数独立

方差分析中的基本假定

1. 在上述假定条件下，判断行业对投诉次数是否有显著影响，实际上也就是检验具有同方差的4个正态总体的均值是否相等
2. 如果4个总体的均值相等，可以期望4个样本的均值也会很接近
   - 4个样本的均值越接近，推断4个总体均值相等的证据也就越充分
   - 样本均值越不同，推断总体均值不同的证据就越充分

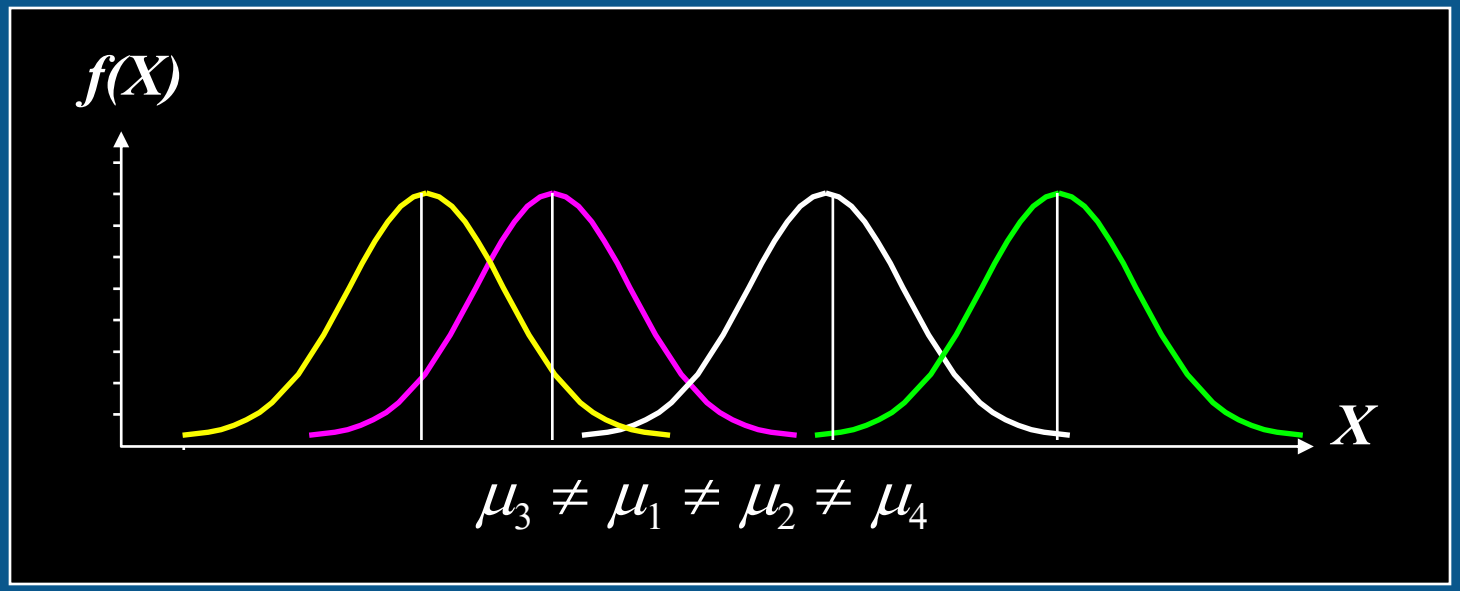
方差分析中基本假定

- 如果原假设成立，即H_0： $\mu_1 = \mu_2 = \mu_3 = \mu_4$
  - 4个行业被投诉次数的均值都相等
  - 意味着每个样本都来自均值为 μ、方差为σ^2的同一正态总体



方差分析中基本假定

- 若备择假设成立，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
  - 4个样本分别来自均值不同的4个正态总体



问题的一般提法

问题的一般提法

1. 设因素有k个水平，每个水平的均值分别用μ_1，μ_2，…，μ_k表示
2. 要检验k个水平(总体)的均值是否相等，需要提出如下假设：
   - H_0：$\mu_1 = \mu_2 = \ldots = \mu_k$
   - H_1：μ_1，μ_2，…，μ_k不全相等
3. 设μ_1为零售业被投诉次数的均值，μ_2为旅游业被投诉次数的均值，μ_3为航空公司被投诉次数的均值，μ_4为家电制造业被投诉次数的均值，提出的假设为
   - H_0：$\mu_1 = \mu_2 = \mu_3 = \mu_4$
   - H_1：μ_1，μ_2，μ_3，μ_4不全相等

10.2 单因素方差分析

10.2.1 数据结构
10.2.2 分析步骤
10.2.3 关系强度的测量
10.2.4 方差分析中的多重比较

作者：贾俊平，中国人民大学统计学院

单因素方差分析的数据结构
(one-way analysis of variance)

观察值（j）	因素(A) i			
	水平A_1	水平A_2	…	水平A_k
1	x_{11}	x_{21}	…	x_{k1}
2	x_{12}	x_{22}	…	x_{k2}
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	…	x_{kn}

分析步骤

- 提出假设
- 构造检验统计量
- 统计决策

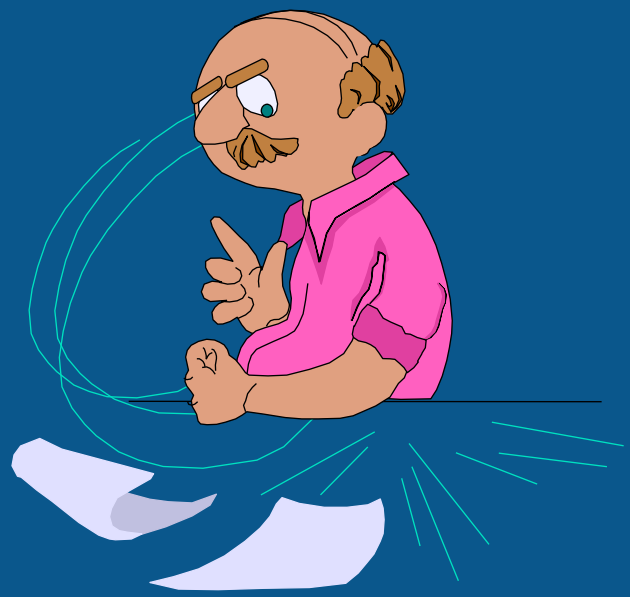
提出假设

1. 一般提法
   - H_0： $\mu_1 = \mu_2 = \ldots = \mu_k$
     - 自变量对因变量没有显著影响
   - H_1： μ_1， μ_2， …， μ_k不全相等
     - 自变量对因变量有显著影响
2. 注意：拒绝原假设，只表明至少有两个总体的均值不相等，并不意味着所有的均值都不相等

构造检验的统计量

构造统计量需要计算

- 水平的均值
- 全部观察值的总均值
- 误差平方和
- 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
(例题分析)

	A	B	C	D	E
1		行业			
2	观测值	零售业	旅游业	航空公司	家电制造业
3	1	57	68	31	44
4	2	66	39	49	51
5	3	49	29	21	65
6	4	40	45	34	77
7	5	34	56	40	58
8	6	53	51		
9	7	44			
10	样本均值	$\bar{x}_1 = 49$	$\bar{x}_2 = 48$	$\bar{x}_3 = 35$	$\bar{x}_4 = 59$
11	样本容量（n）	7	6	5	5
12	总均值	$\bar{\bar{x}} = \dfrac{57+66+\cdots+77+58}{23} = 47.869565$			

构造检验的统计量
(计算总误差平方和 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

  $SST = (57-47.869565)^2+\ldots+(58-47.869565)^2$

  $=115.9295$

构造检验的统计量
(计算组间平方和 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
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 = 1456.608696

构造检验的统计量
(计算组内平方和 *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
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* = 2708

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

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

SST = SSA + SSE

  - 前例的计算结果

4164.608696=1456.608696+2708

构造检验的统计量 (计算均方*MS*)

1. 各误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误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. 由误差平方和除以相应的自由度求得
3. 三个平方和对应的自由度分别是
   - ***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. 组间方差：*SSA*的均方，记为*MSA*，计算公式为

$$MSA=\frac{SSA}{k-1}$$ 前例计算结果：$MSA=\frac{1456.608696}{4-1}=485.536232$

2. 组内方差：*SSE*的均方，记为*MSE*，计算公式为

$$MSE=\frac{SSE}{n-k}$$ 前例计算结果：$MSE=\frac{2708}{23-4}=142.526316$

构造检验的统计量
(计算检验统计量 *F*)

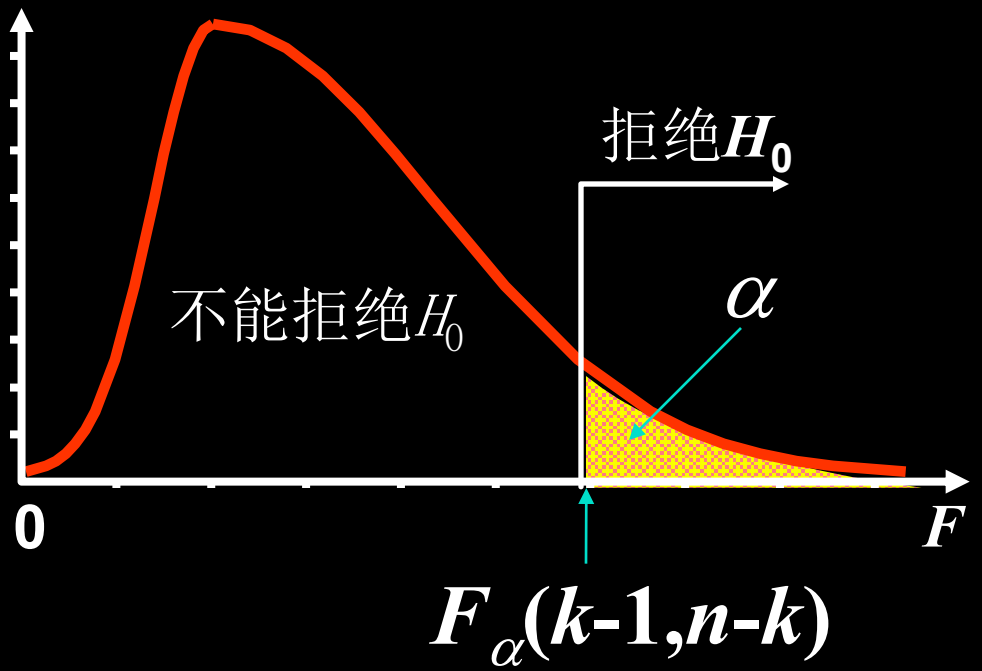
1. 将*MSA*和*MSE*进行对比，即得到所需要的检验统计量*F*
2. 当H_0为真时，二者的比值服从分子自由度为k-1、分母自由度为 n-k 的 *F* 分布，即

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

前例计算结果 $F = \dfrac{485.536232}{142.526316} = 3.406643$

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α的临界值F_α进行比较，作出对原假设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，所检验的因素对观察值有显著影响
  - 若$F<F_\alpha$，则不拒绝原假设H_0，无证据表明所检验的因素对观察值有显著影响

单因素方差分析表
(基本结构)

误差来源	平方和 (SS)	自由度 (df)	均方 (MS)	F值	P值	F 临界值
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
(例题分析)

	A	B	C	D	E	F	G
1	方差分析						
2	差异源	SS	df	MS	F	P-value	F crit
3	组间	1456.608696	3	485.536232	3.406643	0.0387645	3.1273544
4	组内	2708	19	142.526316			
5							
6	总计	4164.608696	22				

用Excel进行方差分析
(Excel分析步骤)

第1步：选择“工具 ”下拉菜单

第2步：选择【数据分析】选项

第3步：在分析工具中选择【单因素方差分析】，然后选择【确定】

第4步：当对话框出现时

在【输入区域 】方框内键入数据单元格区域

在【α】方框内键入0.05(可根据需要确定)

在【输出选项 】中选择输出区域

关系强度的测量

关系强度的测量

1. 拒绝原假设表明因素(自变量)与观测值之间有显著关系
2. 组间平方和(*SSA*)度量了自变量(行业)对因变量(投诉次数)的影响效应
   - 只要组间平方和*SSA*不等于0，就表明两个变量之间有关系(只是是否显著的问题)
   - 当组间平方和比组内平方和(*SSE*)大，而且大到一定程度时，就意味着两个变量之间的关系显著，大得越多，表明它们之间的关系就越强。反之，就意味着两个变量之间的关系不显著，小得越多，表明它们之间的关系就越弱

关系强度的测量

1. 变量间关系的强度用自变量平方和(*SSA*)占总平方和(*SST*)的比例大小来反映
2. 自变量平方和占总平方和的比例记为R^2 ,即

$$R^2 = \frac{SSA(\text{组间平方和})}{SST(\text{总平方和})}$$

3. 其平方根*R*就可以用来测量两个变量之间的关系强度

关系强度的测量
(例题分析)

$$R^2 = \frac{SSA}{SST} = \frac{1456.608696}{4146.608696} = 0.349759 = 34.9759\%$$

R=0.591404

结论

- 行业(自变量)对投诉次数(因变量)的影响效应占总效应的34.9759%，而残差效应则占65.0241%。即行业对投诉次数差异解释的比例达到近35%，而其他因素(残差变量)所解释的比例近为65%以上
- *R*=0.591404，表明行业与投诉次数之间有中等以上的关系

方差分析中的多重比较
(multiple comparison procedures)

作者：贾俊平，中国人民大学统计学院

多重比较的意义

1. 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. 可采用Fisher提出的最小显著差异方法，简写为***LSD***
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

多重比较的步骤

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. 计算检验的统计量: $\bar{x}_i-\bar{x}_j$
3. 计算*LSD*

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

4. 决策：若$\left|\bar{x}_i-\bar{x}_j\right|>LSD$，拒绝$H_0$；若$\left|\bar{x}_i-\bar{x}_j\right|<LSD$，不拒绝$H_0$

多重比较分析
(例题分析)

第1步：提出假设

- 检验1：H_0：$\mu_1 = \mu_2$，H_1：$\mu_1 \neq \mu_2$
- 检验2：H_0：$\mu_1 = \mu_3$，H_1：$\mu_1 \neq \mu_3$
- 检验3：H_0：$\mu_1 = \mu_4$，H_1：$\mu_1 \neq \mu_4$
- 检验4：H_0：$\mu_2 = \mu_3$，H_1：$\mu_2 \neq \mu_3$
- 检验5：H_0：$\mu_2 = \mu_4$，H_1：$\mu_2 \neq \mu_4$
- 检验6：H_0：$\mu_3 = \mu_4$，H_1：$\mu_3 \neq \mu_4$

方差分析中的多重比较
(例题分析)

第2步：计算检验统计量

- 检验1：$|\bar{x}_1 - \bar{x}_2| = |49 - 48| = 1$
- 检验2：$|\bar{x}_1 - \bar{x}_3| = |49 - 35| = 14$
- 检验3：$|\bar{x}_1 - \bar{x}_4| = |49 - 59| = 10$
- 检验4：$|\bar{x}_2 - \bar{x}_3| = |48 - 35| = 13$
- 检验5：$|\bar{x}_2 - \bar{x}_4| = |48 - 59| = 11$
- 检验6：$|\bar{x}_3 - \bar{x}_4| = |35 - 59| = 24$

方差分析中的多重比较
(例题分析)

第3步：计算*LSD*

- 检验1：$LSD_1 = 2.093 \times \sqrt{142.526316 \times (\frac{1}{7} + \frac{1}{6})} = 13.90$
- 检验2：$LSD_2 = 2.093 \times \sqrt{142.526316 \times (\frac{1}{7} + \frac{1}{5})} = 14.63$
- 检验3：$LSD_3 = LSD_2 = 14.63$
- 检验4：$LSD_4 = 2.093 \times \sqrt{142.526316 \times (\frac{1}{6} + \frac{1}{5})} = 15.13$
- 检验5：$LSD_5 = LSD_4 = 15.13$
- 检验6：$LSD_6 = 2.093 \times \sqrt{142.526316 \times (\frac{1}{5} + \frac{1}{5})} = 15.80$

方差分析中的多重比较
(例题分析)

第4步：作出决策

$|\bar{x}_1 - \bar{x}_2| = 1 < 13.90$ 不能认为零售业与旅游业均值之间有显著差异

$|\bar{x}_1 - \bar{x}_3| = 14 < 14.63$ 不能认为零售业与航空公司均值之间有显著差异

$|\bar{x}_1 - \bar{x}_4| = 10 < 14.63$ 不能认为零售业与家电业均值之间有显著差异

$|\bar{x}_2 - \bar{x}_3| = 13 < 15.13$ 不能认为旅游业与航空业均值之间有显著差异

$|\bar{x}_2 - \bar{x}_4| = 11 < 15.13$ 不能认为旅游业与家电业均值之间有显著差异

$|\bar{x}_3 - \bar{x}_4| = 24 > 15.80$ 航空业与家电业均值*有*显著差异

10.3 双因素方差分析

10.3.1 双因素方差分析及其类型
10.3.2 无交互作用的双因素方差分析
10.3.3 有交互作用的双因素方差分析

作者：贾俊平，中国人民大学统计学院

双因素方差分析
(two-way analysis of variance)

1. 分析两个因素(行因素Row和列因素Column)对试验结果的影响
2. 如果两个因素对试验结果的影响是相互独立的，分别判断行因素和列因素对试验数据的影响，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*或*无重复双因素方差分析*(Two-factor without replication)
3. 如果除了行因素和列因素对试验数据的单独影响外，两个因素的搭配还会对结果产生一种新的影响，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*或*可重复双因素方差分析* (Two-factor with replication)

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

无交互作用的双因素方差分析
(无重复双因素分析)

作者：贾俊平，中国人民大学统计学院

双因素方差分析
(例题分析)

【例】有4个品牌的彩电在5个地区销售，为分析彩电的品牌(品牌因素)和销售地区(地区因素)对销售量的影响，对每显著个品牌在各地区的销售量取得以下数据。试分析品牌和销售地区对彩电的销售量是否有显著影响？(α=0.05)

不同品牌的彩电在5个地区的销售量数据

品牌因素	地区因素				
	地区1	地区2	地区3	地区4	地区5
品牌1	365	350	343	340	323
品牌2	345	368	363	330	333
品牌3	358	323	353	343	308
品牌4	288	280	298	260	298

数据结构

	A	B	C	D	E	F	G
1			列因素 (j)				平均值
2			列1	列2	…	列r	$\bar{x}_{i\cdot}$
3	行因素 (i)	行1	x_{11}	x_{12}	…	x_{1r}	$\bar{x}_{1\cdot}$
4		行2	x_{21}	x_{22}	…	x_{2r}	$\bar{x}_{2\cdot}$
5		⋮	⋮	⋮	⋮	⋮	⋮
6		行k	x_{k1}	x_{k2}	…	x_{kr}	$\bar{x}_{k\cdot}$
7	平均值 $\bar{x}_{\cdot j}$		$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	…	$\bar{x}_{\cdot r}$	$\bar{\bar{x}}$
8							

数据结构

- $\bar{x}_{i.}$是行因素的第i个水平下各观察值的平均值

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

- $\bar{x}_{.j}$是列因素的第j个水平下各观察值的平均值

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

- $\bar{\bar{x}}$是全部 kr 个样本数据的总平均值

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

分析步骤
(提出假设)

- 提出假设
  - 对行因素提出的假设为
    - H_0：$\mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_k$ (μ_i为第i个水平的均值)
    - H_1：μ_i (i =1,2, … , k) 不全相等
  - 对列因素提出的假设为
    - H_0：$\mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_r$ (μ_j为第j个水平的均值)
    - H_1：μ_j (j =1,2,…,r) 不全相等

分析步骤
(构造检验的统计量)

- 计算平方和(SS)
  - 总误差平方和 $SST=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{\bar{x}}\right)^2$
  - 行因素误差平方和 $SSR=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{i.}-\bar{\bar{x}}\right)^2$
  - 列因素误差平方和 $SSC=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{.j}-\bar{\bar{x}}\right)^2$
  - 随机误差项平方和 $SSE=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}\right)^2$

分析步骤
(构造检验的统计量)

- 总误差平方和(*SST*)、行因素平方和 (*SSR*)、列因素平方和(*SSC*) 、误差项平方和(*SSE*)之间的关系

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

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

$$SST = SSR + SSC + SSE$$

分析步骤
(构造检验的统计量)

- 计算均方(*MS*)
  - 误差平方和除以相应的自由度
  - 三个平方和的自由度分别是
    - 总误差平方和*SST*的自由度为 ***kr*-1**
    - 行因素平方和*SSR*的自由度为 ***k*-1**
    - 列因素平方和*SSC*的自由度为 ***r*-1**
    - 误差项平方和*SSE*的自由度为 **(*k*-1)×(*r*-1)**

分析步骤
(构造检验的统计量)

- 计算均方(*MS*)
  - 行因素的均方，记为*MSR*，计算公式为

$$MSR = \frac{SSR}{k-1}$$

  - 列因素的均方，记为*MSC*，计算公式为

$$MSC = \frac{SSC}{r-1}$$

  - 误差项的均方，记为*MSE*，计算公式为

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

分析步骤
(构造检验的统计量)

- 计算检验统计量(F)
  - 检验行因素的统计量

$$F_{\mathrm{R}} = \frac{MSR}{MSE} \sim F\left(k-1,(k-1)(r-1)\right)$$

  - 检验列因素的统计量

$$F_{\mathrm{C}} = \frac{MSC}{MSE} \sim F\left(r-1,(k-1)(r-1)\right)$$

分析步骤
(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，作出对原假设H_0的决策
  - 根据给定的显著性水平α在F分布表中查找相应的临界值 F_α
  - 若$F_R > F_\alpha$，拒绝原假设H_0，表明均值之间的差异是显著的，即所检验的行因素对观察值有显著影响
  - 若$F_C > F_\alpha$，拒绝原假设H_0，表明均值之间有显著差异，即所检验的列因素对观察值有显著影响

双因素方差分析表
(基本结构)

误差来源	平方和 (SS)	自由度 (df)	均方 (MS)	F值	P值	F 临界值
行因素	***SSR***	***k*-1**	***MSR***	$\frac{MSR}{MSE}$		
列因素	***SSC***	***r*-1**	***MSC***	$\frac{MSC}{MSE}$		
误差	***SSE***	**(*k*-1)(*r*-1)**	***MSE***			
总和	***SST***	***kr*-1**				

双因素方差分析
(例题分析)

- 提出假设
  - 对品牌因素提出的假设为
    - H_0：$\mu_1=\mu_2=\mu_3=\mu_4$　　(品牌对销售量无显著影响)
    - H_1：μ_i (i =1,2, … , 4) 不全相等　(有显著影响)
  - 对地区因素提出的假设为
    - H_0：$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$　(地区对销售量无显著影响)
    - H_1：μ_j (j =1,2,…,5) 不全相等　(有显著影响)

双因素方差分析
(例题分析)

差异源	SS	df	MS	F	P-value	F crit
行(品牌)	13004.6	3	4334.85	18.1078	9.46E-05	3.4903
列(地区)	2011.7	4	502.925	2.10085	0.14367	3.2592
误差	2872.7	12	239.392			
总和	17889	19				

结论：

- F_R＝18.10777>F_α＝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_C＝2.100846< F_α＝3.2592，不拒绝原假设H_0，无证据表明销售地区对彩电的销售量有显著影响

双因素方差分析
(关系强度的测量)

1. 行平方和(*SSR*)度量了品牌这个自变量对因变量(销售量)的影响效应
2. 列平方和(*SSC*)度量了地区这个自变量对因变量(销售量)的影响效应
3. 这两个平方和加在一起则度量了两个自变量对因变量的联合效应
4. 联合效应与总平方和的比值定义为R^2

$$R^2 = \frac{\text{联合效应}}{\text{总效应}} = \frac{SSR + SSC}{SST}$$

5. 其平方根*R*反映了这两个自变量合起来与因变量之间的关系强度

双因素方差分析
(关系强度的测量)

- 例题分析

$$R^2 = \frac{SSR + SSC}{SST} = \frac{13004.55 + 2011.70}{17888.95} = 0.8394 = 83.94\%$$

  - 品牌因素和地区因素合起来总共解释了销售量差异的83.94%
  - 其他因素(残差变量)只解释了销售量差异的16.06%
  - *R*=0.9162，表明品牌和地区两个因素合起来与销售量之间有较强的关系

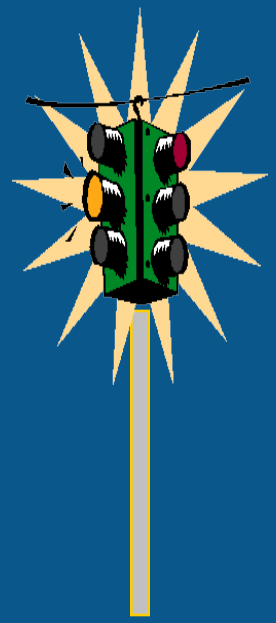
有交互作用的双因素方差分析
(可重复双因素分析)

作者：贾俊平，中国人民大学统计学院

可重复双因素分析
(例题)

【例】城市道路交通管理部门为研究不同的路段和不同的时间段对行车时间的影响，让一名交通警察分别在两个路段和高峰期与非高峰期亲自驾车进行试验，通过试验共获得20个行车时间(单位：min)的数据，如下表。试分析路段、时段以及路段和时段的交互作用对行车时间的影响

	A	B	C	D
1			路段（列变量）	
2			路段1	路段2
3	行车时间(分钟)	高峰期	26	19
4			24	20
5			27	23
6			25	22
7			25	21
8		非高峰期	20	18
9			17	17
10			22	13
11			21	16
12			17	12



可重复双因素方差分析表
(基本结构)

误差来源	平方和 (SS)	自由度 (df)	均方 (MS)	F值	P值	F 临界值
行因素	***SSR***	***k*-1**	***MSR***	F_R		
列因素	***SSC***	***r*-1**	***MSC***	F_C		
交互作用	***SSRC***	**(*k*-1)(*r*-1)**	***MSRC***	F_{RC}		
误差	***SSE***	***Kr*(*m*-1)**	***MSE***			
总和	***SST***	***n*-1**			m为样本的行数	

可重复双因素分析
(平方和的计算)

设：x_{ijl} 为对应于行因素的第i个水平和列因素的第j个水平的第l行的观察值

$\bar{x}_{i.}$ 为行因素的第i个水平的样本均值

$\bar{x}_{.j}$ 为列因素的第j个水平的样本均值

$\bar{x}_{ij}$ 对应于行因素的第i个水平和列因素的第j个水平组合的样本均值

$\bar{\bar{x}}$ 为全部n个观察值的总均值

可重复双因素分析
(平方和的计算)

1. 总平方和：$SST = \sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m}(x_{ijl} - \bar{\bar{x}})^2$

2. 行变量平方和：$SSR = rm\sum_{i=1}^{k}(\bar{x}_{i.} - \bar{\bar{x}})^2$

3. 列变量平方和：$SSC = km\sum_{j=1}^{r}(\bar{x}_{.j} - \bar{\bar{x}})^2$

4. 交互作用平方和：$SSRC = m\sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2$

5. 误差项平方和：$SSE = SST - SSR - SSC - SSRC$

SST=SSR+SSC+SSRC+SSE

可重复双因素分析
(Excel检验步骤)

第1步：选择“工具”下拉菜单，并选择【数据分析】选项

第2步：在分析工具中选择【方差分析：可重复双因素分析】，然后选择【确定】

第3步：当对话框出现时

在【输入区域】方框内键入数据区域(A1：C11)

在【α】方框内键入0.05(可根据需要确定)

在【每一样本的行数】方框内键入重复试验次数(5)

在【输出区域】中选择输出区域

选择【确定】

本章小结

1. 方差分析(ANOVA)的概念
2. 方差分析的思想和原理
3. 方差分析中的基本假设
4. 单因素方差分析
5. 双因素方差分析

结　束

THANKS