方差分析

主讲人：刘宏志

liuhz@ss.pku.edu.cn

北京大学

方差分析简介

- 方差分析，简称ANOVA (Analysis of Variance)
- R.A. Fisher于1923年提出的一种统计方法
- 为纪念Fisher，以*F*命名，故方差分析又称*F*检验
- 检验多个总体的均值是否存在显著差异
- 应用条件：
  - 各样本都来自正态总体
  - 各个总体方差相等（方差齐同）
  - 各样本是相互独立的随机样本

引例：测试营销

- 测试营销常常被用于评估营销组合中的一个或多个要素变动时消费者的反应
- 营销经理们通过试验来确定在不同的广告策略下、不同区域，销量是否存在差异
- 而同一广告策略在不同地区的销售情况往往不同，即区域对销售策略产生影响
- 问题：
  - 不同区域对销量是否有影响？
  - 不同广告策略对销量是否有影响？
  - 广告策略是否会受到区域的影响吗？
  - 广告策略与区域是否会产生交互作用？

示例：苹果汁营销

- 某苹果汁厂家开发了一种新的苹果汁
- 为宣传这种新产品，推广策略将在北上广深4个城市中进行展开
- 为确定各城市的销售数量是否有显著的差异，统计了以下数据

周销售量	城市			
	北京	上海	广州	深圳
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			
城市平均销量	49	48	35	59
总平均			47.9	

方差分析问题的提出

- 若各城市的均值全都相等，则意味着“城市”对新产品的销售量没有影响，即它们之间的销售数量没有显著差异

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

- 若各城市的均值不全相等，则意味着“城市”对销售产量是有影响的，它们之间的销售数量有显著差异

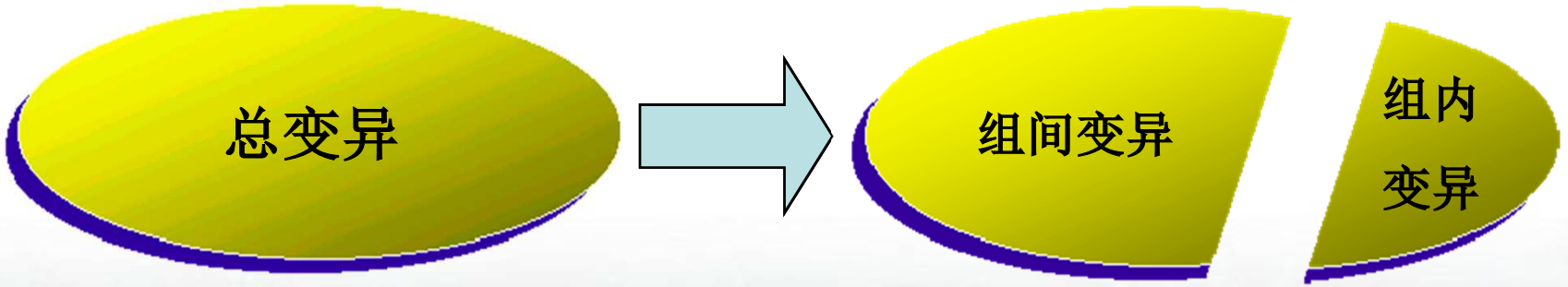
$$H_1: \mu_1, \mu_2, \cdots, \mu_k \text{不全相等}$$

方差分析中的基本概念

- 因素或因子(Factor)：所要检验的对象
  - 例如：要分析城市对推广策略是否有影响，城市是要检验的因素或因子
- 水平或处理(Treatment)：因子的不同表现
  - 例如：北京、广州、上海、深圳就是因子的水平
- 观察值：在每个因素水平下得到的样本数据
  - 例如：每个城市的周销售量就是观察值

方差分析的基本原理和方法

- 样本数据波动的来源：
  - ➢因素中的不同水平
  - ➢抽选样本的随机性
- 波动的度量：
  - ➢组间方差：水平之间的方差
  - ➢组内方差：水平内部的方差



方差分析的基本原理和方法

- 组间方差反映不同因子对样本波动的影响
- 组内方差则是不考虑组间方差的纯随机影响
- 若不同水平对结果无影响，则组间方差中仅有随机因素的差异，而无系统性的差异，它与组内方差就应该近似，两个方差比值应接近于1
- 反之，两个方差的比值就会显著地大于1，当这个比值大到某个程度，或者说达到某临界点，就可以判断出不同水平之间存在着显著性差异
- 小概率原理是方差分析的指导思想

组内方差仅包括随机性因素
组间方差既包括系统性因素，也包括随机性因素

方差分析的基本原理和方法

- 产生方差的独立变量的个数对方差大小有影响
  - 独立变量个数越多，方差就有可能越大
  - 独立变量个数越少，方差就有可能越小
- 使用均方差来消除独立变量个数对方差的影响
  - 均方差（Mean Square）等于方差除以独立变量个数
- 引起方差的独立变量的个数，称为自由度

方差分析的基本原理和方法

检验因子影响是否显著通常用如下F统计量：

$$F = \frac{\text{组间均方差}}{\text{组内均方差}}$$

- F统计量越大，越说明组间方差是主要方差来源，因子影响越显著
- F越小，越说明组内方差（随机方差）是主要的方差来源，因子的影响越不显著

方差分析中的前提条件

- 各样本（观察值）是相互独立的
  - 如每个城市的销量与其他城市的销量无关
- 每个总体都应服从正态分布
  - 对于因素的每一水平，观察值来自服从正态分布的总体
  - 比如，每个城市的销量必需服从正态分布
- 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，四个城市销量的方差都相等

单因素方差分析的数据结构

观测值(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}

单因素条件下离差平方和的分解

- 总离差平方和 $SS_T=SS_E+SS_A$
- 总离差平方和SS_T反映了离差平方和的总体情况

$$SS_T=\sum\sum(X_{ij}-\overline{X})^2$$

- 误差项离差平方和SS_E反映的是水平内部，或组内观察值的离散状况

$$SS_E=\sum\sum(X_{ij}-\overline{X}_{i.})^2$$

- 水平项离差平方和SSA反映的是组间差异

$$SS_A=\sum\sum(\overline{X}_{i.}-\overline{X})^2=\sum n\cdot(\overline{X}_{i.}-\overline{X})^2$$

因素作用显著性的检验

- SS_T的自由度为nk-1　　$SS_T = \sum\sum (X_{ij} - \overline{X})^2$
  - 方差是由于变量波动引起的，但所有的nk个变量并不独立，它们满足一个约束条件，真正独立的变量只有nk-1个
- SS_A的自由度为k-1
  - SS_A是因子在不同水平上的均值变化而产生的方差，但k个均值并不独立，它们满足一个约束条件，因此也丢失一个自由度
- SS_E的自由度为nk-k
  - SS_E是由所有在各因素水平上围绕均值的波动产生，它们满足的约束条件一共k个，失去了k个自由度
- $SS_E SS_T$、SS_A和SS_E的自由度满足如下关系

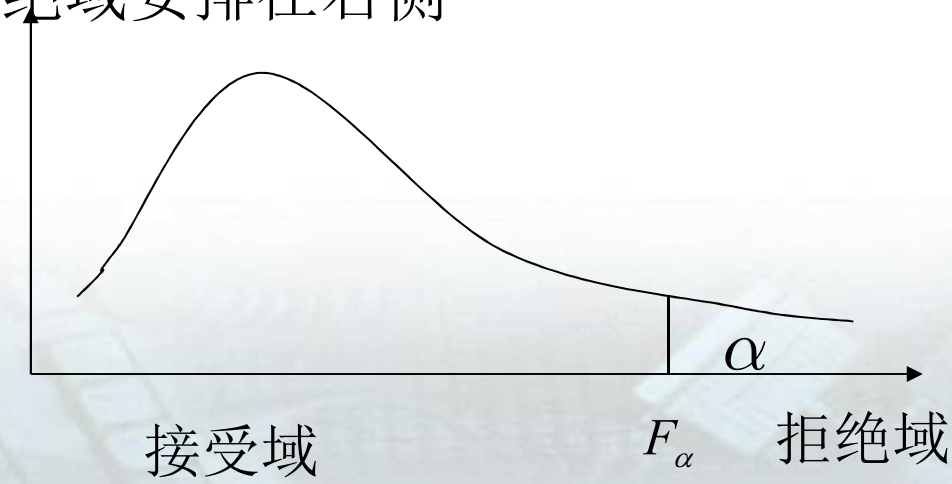
$$nk\text{-}1=(k\text{-}1)+(nk\text{-}k)$$

因素作用显著性的检验

检验统计量： $F = \dfrac{MS_A}{MS_E} \sim F(k-1, n\cdot k-k)$

其中， $MS_A = \dfrac{SS_A}{k-1}$， $MS_E = \dfrac{SS_E}{n\cdot k-k}$

- F值越大，越说明总的方差波动中，组间方差是主要部分，有利于拒绝原假设接受备选假设
- F值越小，越说明随机方差是主要的方差来源，有利于接受原假设，有充分证据说明待检验的因素对总体波动有显著影响
- 检验的拒绝域安排在右侧



方差分析的一般步骤

第1步：提出假设

H_0： $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1： $\mu_1, \mu_2, \ldots,$ μ_k不全相等

方差分析的一般步骤

第2步：构造检验的统计量

观察值的总均值：$\bar{\bar{x}}=\frac{1}{n\cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}=\frac{1}{n\cdot k}\sum_{i=1}^{k}n\cdot\bar{x}_i$

总误差平方和：$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{\bar{x}}\right)^2$

水平项平方和：$SS_A=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\bar{x}_i-\bar{\bar{x}}\right)^2=\sum_{i=1}^{k}n\cdot\left(\bar{x}_i-\bar{\bar{x}}\right)^2$

误差项平方和：$SS_E=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2$

方差分析的一般步骤

组间方差SS_A的均方：　$MS_A = \dfrac{SS_A}{k-1}$

组内方差SS_E的均方：　$MS_E = \dfrac{SS_E}{n \cdot k - k}$

F统计量：　$F = \dfrac{MS_A}{MS_E} \sim F(k-1, n \cdot k - k)$

方差分析的一般步骤

第3步：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

- 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$F<F_\alpha$，则不能拒绝原假设H_0，表明所检验的因素对观察值没有显著影响

方差分析的一般步骤

第4步：根据上述步骤计算的数据，列出方差分析表

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）	SS_A	k-1	MS_A	MS_A/MS_E		
组内（误差）	SS_E	N-k	MS_E			
总和	SS_T	N-1				

N: 观察值总个数，等于nk

示例：苹果汁销售

周销售量	城市			
	北京	上海	广州	深圳
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			
城市平均销量	49	48	35	59
总平均	47.9			

- $SS_T = 4164.609,\quad SS_A = 1456.609,\quad SS_E = 2708$
- $MS_A = 1456.609/(4\text{-}1) = 485.536232,\ MS_E = 2708/(23\text{-}4) = 142.526316$
- $F = MS_A / MS_E = 3.4066643$

示例：苹果汁销售

给定显著性水平$\alpha=0.05$，根据第一自由度$df_1=3$，第二自由度$df_2=19$，查F分布表得到临界值$F_{0.05}(3,19)=3.13$

由于$F > F_{0.05}$，拒绝原假设，即

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

不成立，表明他们之间有显著差异，可以认为城市对销量有显著影响。

示例：苹果汁销售

方差分析表

差异源	SS	df	MS	F	P-value	F crit
组间	1456.609	3	485.536	3.407	0.039	3.127
组内	2708	19	142.526			
总计	4164.609	22				

关系强度的测量

- 拒绝原假设表明因素(自变量)与观测值之间有关系
- 组间平方和(SS_A)度量了自变量(城市)对因变量(销售量)的影响效应：
  - 只要组间平方和SS_A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SS_A比组内平方和SS_E大，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，小得越多，表明它们之间的关系就越弱

关系强度的测量

- 变量间关系的强度可用自变量平方和SS_A 占总平方和SS_T的比例大小来度量
- 自变量平方和占总平方和的比例记为R^2,即

$$R^2 = \frac{SS_A}{SS_T}$$

- 其平方根R 可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：计算可得

$$R^2 = \frac{SS_A}{SS_T} = \frac{1456.608696}{4146.608696} = 0.349759 = 34.9759\%$$

$$R = 0.591404$$

结论：

- 城市(自变量)对销量(因变量)的影响效应占总效应的34.9759%，而残差效应则占65.0241%。即城市对销量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表明城市与销量之间有中等以上的关系

计算机求解（演示）

北京大学

单因素方差分析中的多重比较

- 通过方差分析方法可判断各总体均值是否相等
- 如果不相等，那么如何进一步检验到底哪些均值之间存在差异？
- 多重比较方法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
- 英国统计学家Fisher提出的最小显著差异方法(Least Significant Difference, LSD)是多重比较的常用方法之一
- 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LSD方法

第1步：提出原假设：

$H_0: \mu_i = \mu_j$ (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)

第2步：计算检验的统计量：$\bar{x}_i - \bar{x}_j$

第3步：计算LSD

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

第4步：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

多重比较的LSD方法

第1步：提出假设

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

多重比较的LSD方法

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：$|\bar{x}_3 - \bar{x}_4| = |35 - 59| = 24$

多重比较的LSD方法

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：$LSD_2 = 2.093 \times \sqrt{142.526316 \times (\frac{1}{7} + \frac{1}{5})} = 10.23$
- 检验3：$LSD_3 = LSD_2 = 10.23$
- 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：$LSD_6 = 2.093 \times \sqrt{142.526316 \times (\frac{1}{5} + \frac{1}{5})} = 15.80$

多重比较的LSD方法

第4步：作出决策

$|\bar{x}_1 - \bar{x}_2| = 1 < 13.90$ 北京与上海的销量均值之间*没有*显著差异

$|\bar{x}_1 - \bar{x}_3| = 14 > 10.23$ 北京与广州的销量均值之间*有*显著差异

$|\bar{x}_1 - \bar{x}_4| = 10 < 10.23$ 北京与深圳的销量均值之间*没有*显著差异

$|\bar{x}_2 - \bar{x}_3| = 13 < 15.13$ 上海与广州的销量均值之间*没有*显著差异

$|\bar{x}_2 - \bar{x}_4| = 11 < 15.13$ 上海与深圳的销量均值之间*没有*显著差异

$|\bar{x}_3 - \bar{x}_4| = 24 > 15.80$ 广州与深圳的销量均值之间*有*显著差异

双因素方差分析

北京大学

双因素方差分析

- 动机：
  - 单因素方差分析只考虑一个自变量对数值型因变量的影响
  - 实际中，有时需要考虑两个或多个因素对实验结果的影响
- 当方差分析中涉及两个自变量时，称为双因素方差分析(Two-way Analysis of Variance)
- 获取数据时，将一个因素安排在“行”的位置，称为行因素，另一因素安排在“列”的位置，称为列因素
- 前提：
  - 每个总体都服从正态分布、各总体的方差齐同、各观察值独立
- 类型：
  - 无交互作用的双因素方差分析
  - 有交互作用的双因素方差分析

示例：苹果汁销售

不同广告策略下的新产品在各地区的销售量数据

广告策略	区域因素				
	城市A	城市B	城市C	城市D	城市E
策略1	365	350	343	340	323
策略2	345	368	363	330	333
策略3	358	323	353	343	308
策略4	288	280	298	260	298

假设有4种广告策略分别在5个地区进行推广销售，试分析广告策略和销售区域对新产品的销售量是否有显著影响？

无交互作用的双因素方差分析

		列因素（j）				平均值
		列1	列2	…	列k	$\bar{x}_{i\cdot}$
行因素（i）	行1	x_{11}	x_{12}	…	x_{1k}	$\bar{x}_{1\cdot}$
	行2	x_{21}	x_{22}	…	x_{2k}	$\bar{x}_{2\cdot}$
	⋮	⋮	⋮	⋮	⋮	⋮
	行r	x_{r1}	x_{r2}	…	x_{rk}	$\bar{x}_{r\cdot}$
平均值 $\bar{x}_{\cdot j}$		$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	…	$\bar{x}_{\cdot k}$	$\bar{\bar{x}}$

无交互作用的双因素方差分析

- $\bar{x}_{i.}$ 是行因素的第i个水平下各观察值的平均值

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

- $\bar{x}_{.j}$ 是列因素的第j个水平下各观察值的平均值

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

- $\bar{\bar{x}}$ 是全部 kr 个样本数据的总平均值

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

无交互作用的双因素方差分析

第1步：提出假设

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

无交互作用的双因素方差分析

第2步：构造检验的统计量

计算平方和(SS)

总误差平方和：
$$SS_T=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{\bar{x}}\right)^2$$

行因素误差平方和：
$$SS_R=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{i.}-\bar{\bar{x}}\right)^2$$

列因素误差平方和：
$$SS_C=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{.j}-\bar{\bar{x}}\right)^2$$

随机误差项平方和：
$$SS_E=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}\right)^2$$

无交互作用的双因素方差分析

总离差平方和SS_T、水平项离差平方和SS_R和SS_C、误差项离差平方和SS_E之间的关系

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

$$SS_T = SS_R + SS_C + SS_E$$

无交互作用的双因素方差分析

- 计算均方MS

  误差平方和除以相应的自由度

- 各平方和的自由度
  - 总离差平方和SS_T的自由度为 kr-1
  - 行因素的离差平方和SS_R的自由度为 r-1
  - 列因素的离差平方和SS_C的自由度为 k-1
  - 随机误差平方和SS_E的自由度为$(k\text{-}1)\times(r\text{-}1)$

无交互作用的双因素方差分析

- 计算均方MS
  - 行因素的均方，记为MS_R，计算公式为

$$MS_R = \frac{SS_R}{r-1}$$

  - 列因素的均方，记为MS_C，计算公式为

$$MS_C = \frac{SS_C}{k-1}$$

  - 随机误差项的均方，记为MS_E，计算公式为

$$MS_E = \frac{SS_E}{(k-1)(r-1)}$$

无交互作用的双因素方差分析

- 计算检验统计量F
  - 检验行因素的统计量

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

  - 检验列因素的统计量

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

无交互作用的双因素方差分析

第3步：决策分析

- 根据给定的显著性水平α和计算出的自由度在F分布表中查找相应的临界值F_α
- 将统计量的值F_R、F_C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
  - 若$F_R > F_\alpha$，则拒绝原假设H_0，表明均值之间的差异是显著的，即所检验的行因素对观察值有显著影响
  - 若$F_C > F_\alpha$，则拒绝原假设H_0，表明均值之间有显著差异，即所检验的列因素对观察值有显著影响

无交互作用的双因素方差分析

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	SS_R	k-1	MS_R	MS_R/MS_E		
列因素	SS_C	r-1	MS_C	MS_C/MS_E		
随机因素	SS_E	(k-1)(r-1)	MS_E			
总和	SS_T	kr-1				

示例：苹果汁销售

提出假设：

- 对策略因素提出的假设为

  H_0： $\mu_1=\mu_2=\mu_3=\mu_4$ (策略对销售量没有影响)

  H_1： μ_i (i =1,2,…,4)不全相等 (策略对销售量有影响)

- 对城市因素提出的假设为

  H_0： $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$ (城市对销售量没有影响)

  H_1： μ_j (j =1,2,…,5)不全相等 (城市对销售量有影响)

示例：苹果汁销售

差异源	SS	df	MS	F	P-value	Fcrit
行	13004.55	3	4334.85	18.10777	9.46E-05	3.490295
列	2011.7	4	502.925	2.100846	0.143665	3.259167
误差	2872.7	12	239.3917			
总计	17888.95	19				

结论： F_R=18.10777>F_α=3.4903，拒绝原假设H_0，说明策略对销售量有显著影响； F_C=2.100846< F_α=3.2592，不能拒绝原假设H_0，说明销售城市对销售量没有显著影响

有交互作用的双因素方差分析

- 不仅分析两个因素自身各个独立的效应，还分析两个因素的组合所产生的交互作用进行显著性检验

- 为了进行针对交互作用的显著性检验，需要对两个因素的任意组合进行多次观测，获得多项观测值，因此有交互作用的双因素方差分析又称为可重复双因素方差分析

示例：交通状况分析

城市道路交通管理部门为研究不同路段和不同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

通过试验取得共获得20个行车时间的数据

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

示例：交通状况分析

不同路段不同时间段内20个行车时间数据

		路段（列变量）	
		路段甲	路段乙
时段(行变量)	高峰期	26	19
		24	20
		27	23
		25	22
		25	21
	非高峰期	20	18
		17	17
		22	13
		21	16
		17	12

北京大学

有交互作用的双因素方差分析

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	SS_R	k-1	MS_R	MS_R/MS_E		
列因素	SS_C	r-1	MS_C	MS_C/MS_E		
交互作用	SS_{RC}	$(k-1)(r-1)$	MS_{RC}	MS_{RC}/MS_E		
随机因素	SS_E	$kr(m-1)$	MS_E			
总和	SS_T	n-1				

有交互作用的双因素方差分析

- 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

有交互作用的双因素方差分析

平方和的计算

总平方和：

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m}(x_{ijl} - \bar{\bar{x}})^2$$

行变量平方和：

$$SS_R = rm\sum_{i=1}^{k}(\bar{x}_{i.} - \bar{\bar{x}})^2$$

列变量平方和：

$$SS_C = km\sum_{j=1}^{r}(\bar{x}_{.j} - \bar{\bar{x}})^2$$

交互作用平方和：

$$SS_{RC} = m\sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2$$

误差项平方和：

$$SS_E = SS_T - SS_R - SS_C - SS_{RC}$$

示例：交通状况分析

差异源	SS	df	MS	F	P-value	F crit
样本	174.05	1	174.05	44.06329	5.7E-06	4.493998
列	92.45	1	92.45	23.40506	0.000182	4.493998
交互	0.05	1	0.05	0.012658	0.911819	4.493998
内部	63.2	16	3.95			
总计	329.75	19				

结论：

- 时段(行因素)对行车时间有显著影响；
- 路段(列因素)对行车时间有显著影响；
- 无证据表明时段和路段的交互作用对行车时间有显著影响