LOGO

第一讲-3 数据方差分析

方差分析的目的

- 方差分析是检验三个或三个以上总体均值是否存在显著差异的统计推断方法，是假设检验问题的进一步扩展。

方差分析(ANOVA)

1. 20世纪20年代由英国统计学家Ronald A.Fisher首先提出
2. 检验多个总体均值是否相等
3. 研究分类型自变量对数值型因变量的影响
4. 有单因子方差分析和双因子方差分析

方差分析的思想和原理

- 有关概念
  - 在方差分析中，所要检验的对象称为因素或因子（**Factor**）
  - 因子的不同表现称为水平或处置（**Treatment**）
  - 每个因子水平下得到的样本数据称为观察值。

什么是方差分析?
(例题分析)

【例】确定超市的位置和竞争者的数量对销售额是否有显著影响，获得的年销售额数据(单位：万元)如下表

	A	B	C	D	E	F
1	因子		竞争者数量			
2			0个	1个	2个	3个以上
3	超市位置	商业区	410	380	590	470
4			300	310	480	400
5			450	390	510	390
6		居民小区	250	290	440	430
7			310	350	480	420
8			220	300	500	530
9		写字楼	180	220	290	240
10			290	170	280	270
11			330	250	260	320

水平或处理

样本数据

什么是方差分析?
(例题分析)

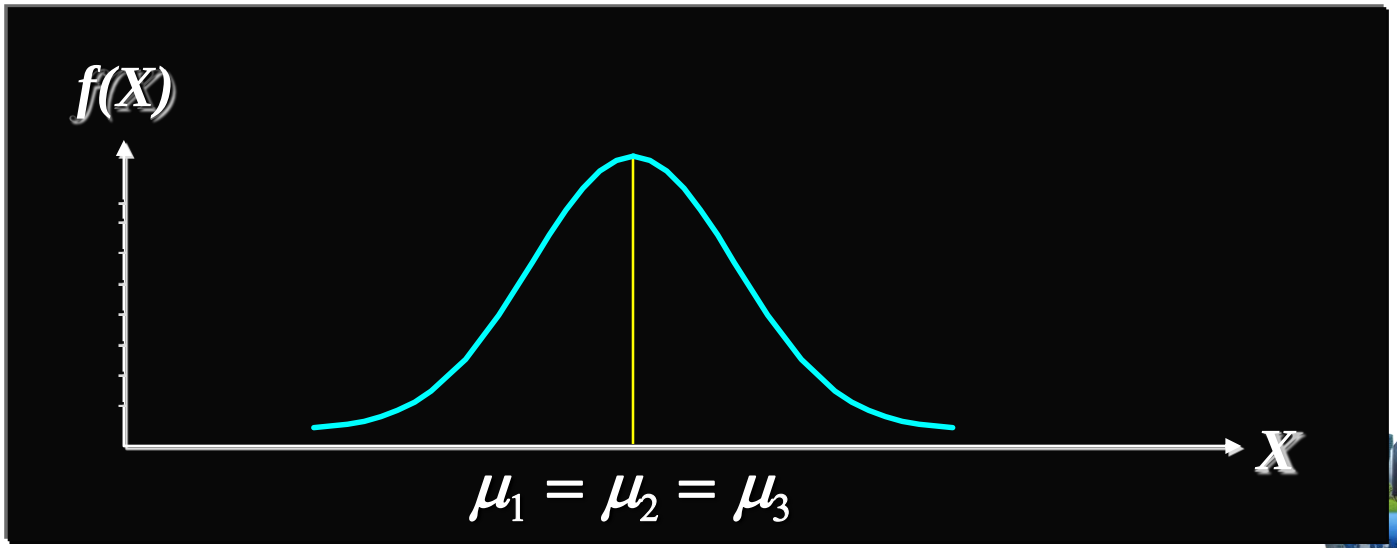
1. 如果只考虑“超市位置”对销售额是否有显著影响，实际上也就是要判断不同位置超市的销售额均值是否相同
   - 若它们的均值相同，意味着“超市位置”对销售额没有显著影响；若均值不全相同，则意味着“超市位置”对销售额有显著影响
   - “超市位置”就是分类自变量，“销售额”则是数值因变量。“超市位置”是要检验的对象，称为因子(factor)，商业区、居民小区、写字楼是因子的3个取值，称为水平(level)或处理(treatment)。每个因子水平下得到的销售额为样本观测值
2. 方差分析要解决的问题就是判断超市的位置对销售额是否有显著影响。设商业区、居民小区和写字楼3个位置超市的销售额均值是否相同

方差分析的基本假定

1. **正态性**。每个总体都应服从正态分布，即对于因子的每一个水平，其观测值是来自正态分布总体的简单随机样本
   - 在上例中，要求每个位置超市的销售额必须服从正态分布
2. **方差齐性**。各个总体的方差必须相同，对于分类变量的个水平，有$\sigma_1^2=\sigma_2^2=\ldots=\sigma_k^2$
   - 在上例中，要求不同位置超市的销售额的方差都相同
3. **独立性**。每个样本数据是来自因子各水平的独立样本(该假定不满足对结果影响较大)
   - 在上例中，3个样本数据是来自不同位置超市的3个独立样本

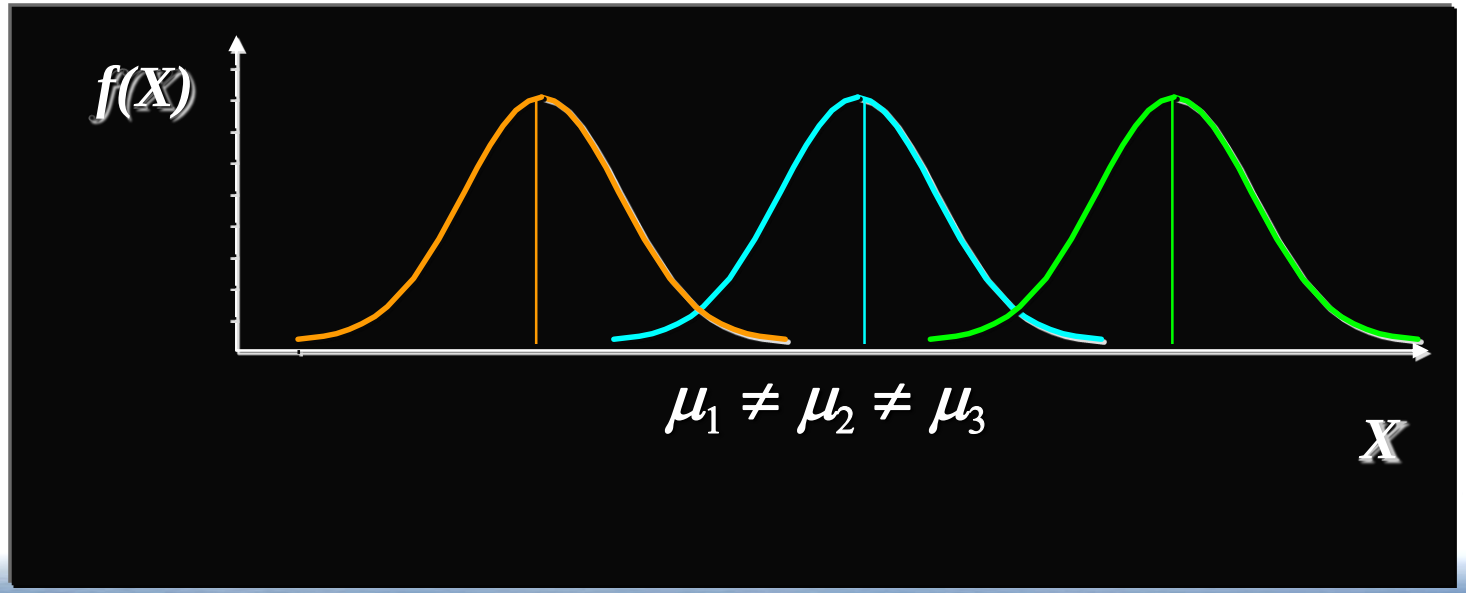
方差分析中基本假定

- 如果原假设成立，即H_0：$\mu_1=\mu_2=\mu_3$
  - 不同位置超市的平均销售额相等
  - 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)不全相等
  - 至少有一个总体的均值是不同的
  - 3个样本分别来自均值不同的3个正态总体



方差分析的基本原理
(误差分解)

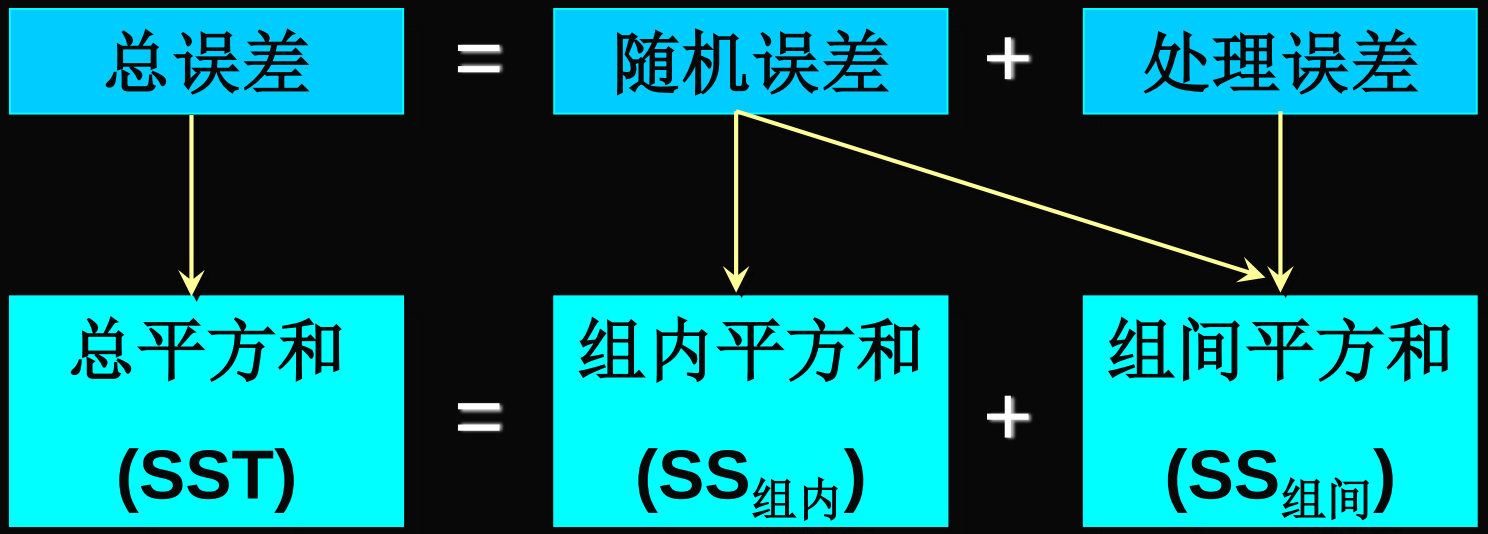
1. 总误差
   - 反映全部观测数据的误差称总误差
   - 所抽取的全部36家超市的销售额之间差异
2. 随机误差—组内误差
   - 由于抽样的随机性造成的误差
   - 反映样本内部数据之间的随机误差
3. 处理误差—组间误差
   - 不同的处理影响所造成的误差
   - 反映样本之间数据的差异

方差分析的基本原理
(误差分解)

1. 数据的误差用平方和(SS)表示
2. 总平方和(SST)
   - 反映全部数据总误差大小的平方和
   - 抽取的全部36家超市销售额之间的误差平方和
3. 组内平方和($SS_{组内}$)
   - 反映组内误差大小的平方和
     - 比如，每个位置超市销售额的误差平方和
   - 只包含随机误差
4. 组间平方和($SS_{组间}$)
   - 反映组间误差大小的平方和
     - 比如，不同位置超市销售额之间的误差平方和
   - 既包括随机误差，也包括处理误差

方差分析的基本原理
(误差分解)

- 误差平方和的分解及其关系



方差分析的基本原理
(误差分析)

1. 判断原假设是否成立，就是判断组间方差与组内方差是否有显著差异
2. 若原假设成立，组间均方误差与组内均方误差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，组间均方误差会大于组内均方误差，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

单因子方差分析

1. 只考虑一个分类型自变量影响的方差分析
   - 比如，在上例中，只考虑超市位置一个因子对销售额度影响，或者只考虑竞争者数量对销售额的影响，都属于单因子方差分析
2. 分析步骤包括

- 提出假设
- 选择显著性水平
- 构造检验的统计量
- 制定决策规则
- 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F

1. 将组间均方和$MS_{组间}$除以组内均方和$MS_{组内}$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

$$F = \frac{MS_{组间}}{MS_{组内}} \sim F(k-1, n-k)$$

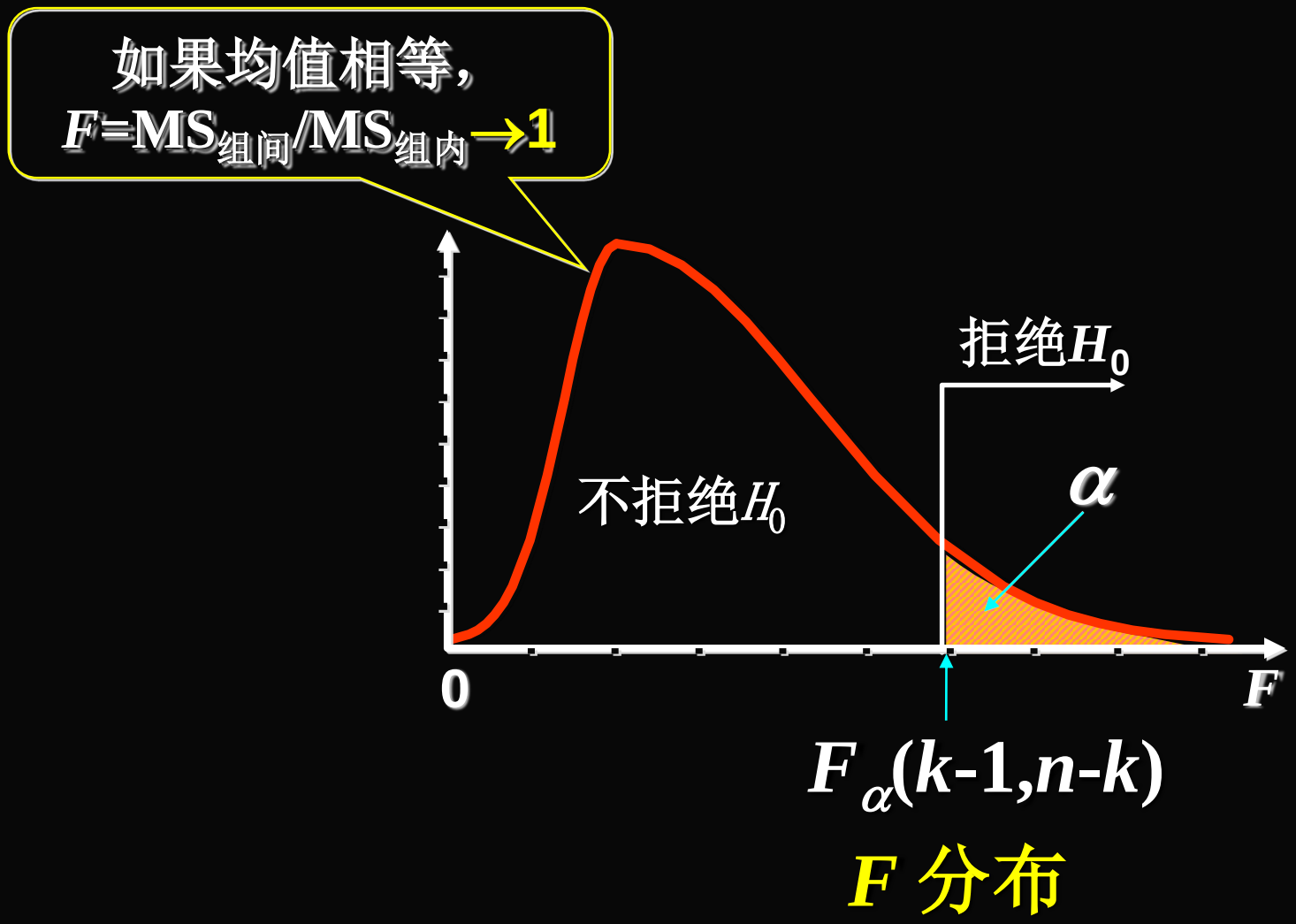
组间平方和 $$SS_{组间} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{\bar{x}})^2$$

组内平方和 $$SS_{组内} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

做出决策

- 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(或计算出统计量的P值)，做出决策
  - 若$F>F_\alpha$，拒绝原假设H_0，表明均值之间的差异是显著的，所检验的因子对观察值有显著影响
  - 若$F<F_\alpha$，不拒绝原假设H_0，无证据表明所检验的因子对观察值有显著影响

作出决策
(F分布与拒绝域)



LOGO

2 单因子方差分析

2.2 哪些均值之间有显著差异？

多重比较的意义

1. 在拒绝原假设的条件下，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. 比较方法有多种，若Fisher提出的最小显著差异方法，简写为*LSD*

多重比较的LSD方法

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: $\left|\bar{x}_i-\bar{x}_j\right|$
3. 计算*LSD*

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

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

LOGO

方差分析

3　双因子方差分析

3.1　不考虑交互作用
3.2　考虑交互作用

LOGO

3 双因子方差分析

3.1 不考虑交互作用

双因子方差分析

1. 分析两个因子(行因子Row和列因子Column)对实验结果的影响
2. 如果两个因子对实验结果的影响是相互独立的，分别判断行因子和列因子对实验数据的影响，这时的双因子方差分析称为*无交互作用的双因子方差分析*或*无重复双因子方差分析*
3. 如果除了行因子和列因子对实验数据的单独影响外，两个因子的搭配还会对结果产生一种新的影响，这时的双因子方差分析称为*有交互作用的双因子方差分析*或*可重复双因子方差分析*

双因子方差分析的基本假定

1. 每个总体都服从正态分布
   - 对于因子的每一个水平，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

双因子方差分析—不考虑交互作用

(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(品牌因子)和销售地区(地区因子)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

品牌因子	地区因子				
	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

分析步骤
(提出假设)

- 提出假设
  - 对行因子提出的假设为
    - 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
  - 对列因子提出的假设为
    - 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

双因子方差分析
(例题分析)

- 提出假设
  - 对品牌因子提出的假设为
    - H_0： $\mu_1=\mu_2=\mu_3=\mu_4$
    - H_1： μ_i (i =1,2, … , 4) 不全相等
  - 对地区因子提出的假设为
    - H_0： $\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$
    - H_1： μ_j (j =1,2,…,5) 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 $SST=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{\bar{x}}\right)^2$
  - 行因子误差平方和 $SSR=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{i.}-\bar{\bar{x}}\right)^2$
  - 列因子误差平方和 $SSC=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(\bar{x}_{.j}-\bar{\bar{x}}\right)^2$
  - 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)、行因子平方和 ($SS_{行}$)、列因子平方和($SS_{列}$) 、误差项平方和($SS_{残差}$)之间的关系

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

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

$$SST = SS_{行} + SS_{列} + SS_{残差}$$

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**
  - 行因子平方和*SSR*的自由度为 ***k*-1**
  - 列因子平方和*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)
  - 行因子的均方，记为$MS_{行}$，计算公式为

$$MS_{行} = \frac{SS_{行}}{k-1}$$

  - 列因子的均方，记为$MS_{列}$，计算公式为

$$MS_{列} = \frac{SS_{列}}{r-1}$$

  - 误差项的均方，记为$MS_{残差}$，计算公式为

$$MS_{残差} = \frac{SS_{残差}}{(k-1)(r-1)}$$

分析步骤
(构造检验的统计量)

- 计算检验统计量(F)
  - 检验行因子的统计量

$$F_R = \frac{\mathrm{MS}_{行}}{\mathrm{MS}_{残差}} \sim F\left(k-1,(k-1)(r-1)\right)$$

  - 检验列因子的统计量

$$F_C = \frac{\mathrm{MS}_{列}}{\mathrm{MS}_{残差}} \sim F\left(r-1,(k-1)(r-1)\right)$$

分析步骤
(做出决策)

- 计算出统计量的P值与给定的显著性水平α比较，
  - 若$P_R<\alpha$，**拒绝**原假设H_0，表明均值之间的差异是显著的，即所检验的行因子对观察值有显著影响
  - 若$P_C<\alpha$，**拒绝**原假设H_0，表明均值之间有显著差异，即所检验的列因子对观察值有显著影响

LOGO

3 双因子方差分析

3.2 考虑交互作用

可重复双因子分析
(提出假设)

- 提出假设
  - 对行因子提出的假设为
    - 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
  - 对列因子提出的假设为
    - 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
  - 对交互作用的假设为
    - H_0：无交互作用
    - H_1：有交互作用

可重复双因子分析
(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：$$SS_{行}=rm\sum_{i=1}^{k}(\bar{x}_{i.}-\bar{\bar{x}})^2$$

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

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

5. 误差项平方和：$$SS_{残差}=SST-SS_{行}-SS_{列}-SSR_{交互}$$

$$SST=SS_{行}+SS_{列}+SS_{交互}+SS_{残差}$$

可重复双因子分析
(构造检验统计量)

1. 检验行因子的统计量

$$F_R = \frac{\text{行因子均方}}{\text{残差均方}} = \frac{\text{MS}_{\text{行}}}{\text{MS}_{\text{残差}}} \sim F(k-1, kr(m-1))$$

2. 检验列因子的统计量

$$F_C = \frac{\text{列因子均方}}{\text{残差均方}} = \frac{\text{MS}_{\text{列}}}{\text{MS}_{\text{残差}}} \sim F(r-1, kr(m-1))$$

3. 检验交互作用的统计量

$$F_{RC} = \frac{\text{交互作用均方}}{\text{残差均方}} = \frac{\text{MS}_{\text{交互}}}{\text{MS}_{\text{残差}}} \sim F((k-1)(r-1), kr(m-1))$$

计算出统计量的P值，若P<α，拒绝原假设

本章小结

- 方差分析的思想和原理
- 方差分析中的基本假设
- 单因子方差分析
- 双因子方差分析

Thank You !


AUSTRALIA
OCEANIA
Northern Territory
Queensland
Western Australia
South Australia
New South Wales
Victoria
Tasmania
NEW ZEALAND
VANUATU
Port-Vila
TONGA
New Caledonia
Chesterfield
Coral Sea
Tasman Sea
Sydney
Canberra ACT
Brisbane
Perth
Darwin
Cairns
Alice Springs
Norfolk I.