双因素方差分析

第三节　双因素方差分析

一、双因素方差分析及其类型

二、无交互作用的双因素方差分析

三、有交互作用的双因素方差分析

双因素方差分析

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

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		$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	…	$\bar{x}_{\cdot r}$	$\bar{\bar{x}}$
8	$\bar{x}_{\cdot j}$						

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）
  - $H_1：\mu_i \ (i = 1, 2, \ldots, k)$ 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）
  - $H_1：\mu_j \ (j = 1, 2, \ldots, r)$ 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 $$SSE=\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{x}_{i.}-\bar{x}_{.j}+\bar{\bar{x}}\right)^2$$

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

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

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

$SST = SSR + SSC + SSE$

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 $(k-1)\times(r-1)$

计算均方(*MS*)

- 行因素的均方，记为*MSR*，计算公式为 $MSR = \dfrac{SSR}{k-1}$

- 列因素的均方，记为*MSC*，计算公式为 $MSC = \dfrac{SSC}{r-1}$

- 误差项的均方，记为*MSE*，计算公式为 $MSE = \dfrac{SSE}{(k-1)(r-1)}$

➡计算检验统计量(*F*)

- 检验行因素的统计量

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

- 检验列因素的统计量

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

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				

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)

差异源	SS	df	MS	F	P-value	F crit
行(品牌)	13004.55	3	4334.85	18.10777	9.46E-05	3.4903
列(地区)	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：

- $F_R = 18.10777 > F_\alpha = 3.4903$，拒绝原假设$H_0$，说明彩电的品牌对销售量有显著影响
- $F_C = 2.100846 < F_\alpha = 3.2592$，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

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

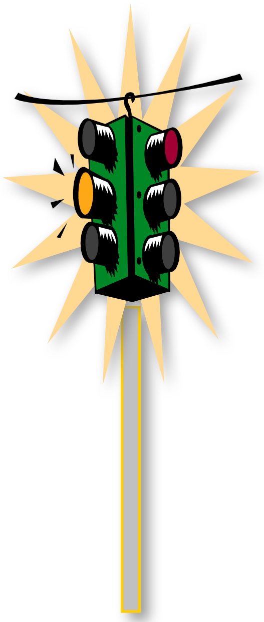
➡例题分析

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

- 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

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	时段(行变量)	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

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

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