方差分析

本章主要内容

- 方差分析基本原理
- 单因素方差分析
- 双因素方差分析

本章主要内容

- 方差分析基本原理
- 单因素方差分析
- 双因素方差分析

方差分析基本原理

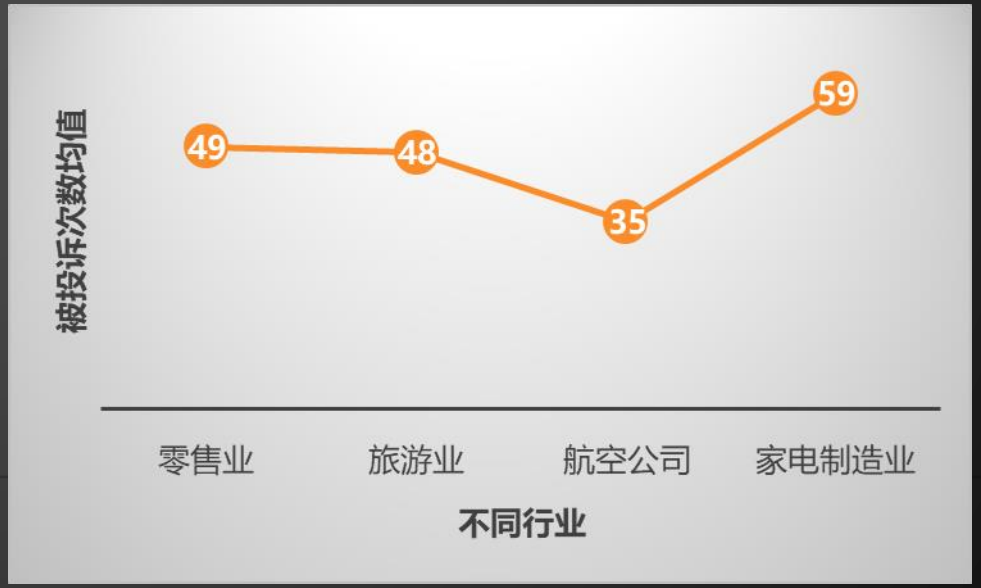
- 方差分析，Analysis of Variance，简称ANOVA，主要用于比较多个总体的均值是否想等。
- 例如，被投诉次数可以反映服务质量，现在有四个行业：零售、旅游业、航空公司、家电制造业，每个行业中抽取若干个企业，其中零售业抽取7家，旅游业抽取6家，航空公司抽取5家，家电制造业抽取5家，统计出它们的被投诉次数，如下表所示。我们想知道这四个行业之间的服务质量是否有显著差异?
- 分析：被投诉次数反映了服务质量，可以用被投诉次数的均值来代表服务质量，所以这四个行业之间的服务质量是否有显著差异就转化为这四个行业之间的被投诉次数的均值是否有显著差异。

行业			
零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			

方差分析基本原理

- 先求出这四个行业的被投诉次数均值，接着绘制出这四个行业被投诉次数均值的折线图，如下图所示。
- 折线图的上下起伏，即波动性越大，说明服务质量差异越大，方差能够衡量波动性的大小，也就说可以通过方差来描述这种差异性。
- 当方差大到一定程度，就说明这四个行业之间的服务质量有显著差异，所以可以通过方差来分析这种问题，称之为方差分析。

行业			
零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			
49	48	35	59

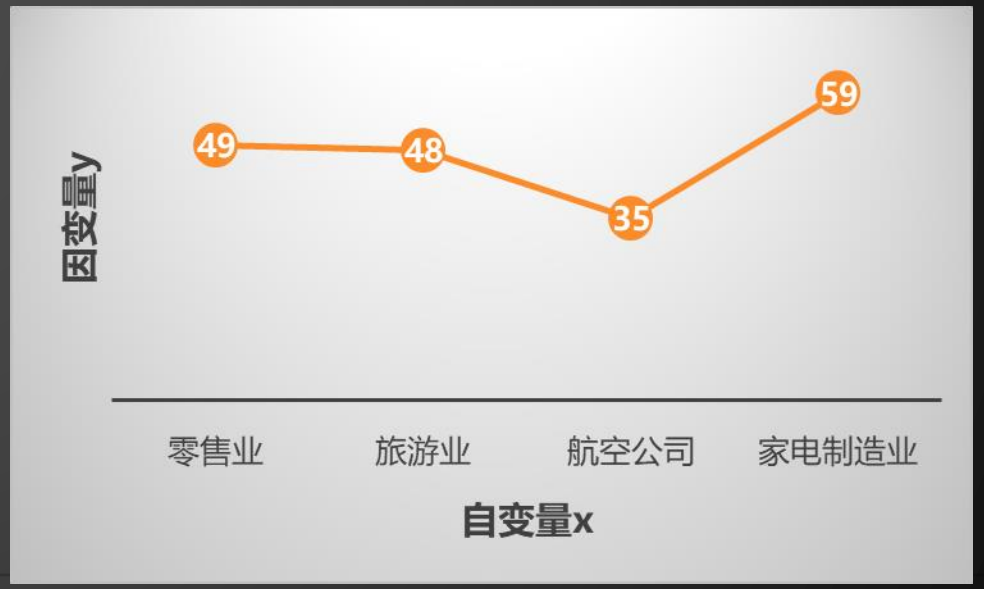


方差分析基本原理

在这个折线图中，横坐标表示的不同行业可以看成是自变量x，纵坐标表示的被投诉次数均值可以看成是因变量y，所以，从形式上看，方差分析是比较多个总体的均值是否相等，但本质上研究的是变量之间的关系，而且是分类型自变量和数值型因变量之间的关系。

应用领域：

- 分析心理学试验数据；
- 分析生物学试验数据；
- 分析工程和医药的试验数据等。



方差分析基本术语

方差分析：通过检验多个总体的均值是否相等来判断分类型自变量对数值型因变量是否有显著影响。

例如，前面的例子中，自变量：行业，因变量：被投诉次数。

相关术语：

- 因素（因子）：检验的对象，行业
- 水平：因素的不同表现，指零售业、旅游业、航空公司、家电制造业
- 观测值：每个因子水平下得到的样本数据，被投诉次数

在服务质量这个例子中，方差分析要研究的就是行业对被投诉次数是否有显著影响。

方差分析中的3个基本假定

1. 每个总体都服从正态分布。例如，在前面的例子中，要求每个行业被投诉次数必须服从正态分布，正态分布就意味着在一个行业中，有些企业的被投诉次数可能很高，有些可能很低，但大多数企业的被投诉次数属于正常情况，而且，确保这些行业都来自于同一分布。
2. 各个总体的方差σ^2必须相同。例如，在前面的例子中，要求每个行业被投诉次数的方差都相同，类似于控制变量，在方差相同的情况，比较这些正态分布的均值是否相等。
3. 观测值是独立的。例如，在前面的例子中，要求每个被抽中企业的被投诉次数都与其他企业的被投诉次数相互独立，即某个行业的被投诉次数只与自身行业相关，与其他行业无关。

方差分析的基本思想

1、图形描述

不同行业被投诉的次数是有明显差异的，而且，在同一个行业，不同企业被投诉的次数也明显不同

方差分析的基本思想

2、误差分析

在方差分析中，误差通过平方和来表述。

- 组内误差：反应组内误差大小的平方和称为组内平方和，记为SSE（英文缩写）。
- 组间误差：反应组间误差大小的平方和称为组间平方和，记为SSA。

组间误差来源：随机误差+行业导致的误差，组内误差来源：随机误差。

如果不同行业对被投诉次数没有影响，即没有行业导致的误差，那么组间误差与组内误差经过平均后的数值（方差或者均方）就应该很接近，即它们的比值就会接近1。反之，它们的比值就会大于1。

- 总误差：反应全部数据误差大小的平方和为总平方和，记为SST。

总平方和（SST）=组内平方和（SSE）+组间平方和（SSA）

本章主要内容

- 方差分析基本原理
- 单因素方差分析
- 双因素方差分析

单因素方差分析

主要内容：

- 单因素方差分析及其数据结构
- 单因素方差分析3步法
- 单因素方差分析效果评估
- 用Excel做单因素方差分析
- 方差分析中的多重比较：LSD检验法

单因素方差分析

主要内容：

- 单因素方差分析及其数据结构
- 单因素方差分析3步法
- 单因素方差分析效果评估
- 用Excel做单因素方差分析
- 方差分析中的多重比较：LSD检验法

单因素方差分析：数据结构

- 当方差分析中只涉及一个分类型自变量时，称为单因素方差分析（one-way analysis of variance），单因素方差分析研究的是一个分类型自变量对一个数值型因变量的影响。
- 例如，前面的例子中，研究行业（分类型自变量）对被投诉次数（数值型因变量）的影响。

行业			
零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			

抽象

观察值	因素			
1	水平A_1	水平A_2	…	水平A_k
2	x_{11}	x_{12}	…	x_{1k}
3	x_{21}	x_{22}		x_{2k}
…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nk}

单因素方差分析

主要内容：

- 单因素方差分析及其数据结构
- 单因素方差分析3步法
- 单因素方差分析效果评估
- 用Excel做单因素方差分析
- 方差分析中的多重比较：LSD检验法

单因素方差分析基本步骤

1、提出假设

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$　　分类型自变量对数值型因变量没有显著影响

$H_1: \mu_1, \mu_2, \cdots, \mu_k$不全相等　　有显著影响

2、构造检验的统计量

误差平方和——>方差——>方差比——>F分布

计算出F值

3、统计决策

根据显著性水平α计算出临界值，和第2步计算出来的F值比较

根据比较结果，作出决策，拒绝还是接受原假设

单因素方差分析：以各行业服务质量为例说明

1、提出假设

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$　　行业对被投诉次数没有显著影响

$H_1: \mu_1, \mu_2, \mu_3, \mu_4$不全相等　　行业对被投诉次数有显著影响

2、构造检验的统计量

- 计算各行业的被投诉次数均值

例如，根据表中的数据，计算零售行业的样本均值为：

$$\bar{x}_1 = \frac{\sum_{j=1}^{7} x_{1j}}{n_1} = \frac{57 + 66 + 49 + 40 + 34 + 53 + 44}{7} = 49$$

行业			
零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			

单因素方差分析：以各行业服务质量为例说明

- 按照同样的方式，可以得到旅游业、航空公司、家电制造业的均值。

	零售业	旅游业	航空公司	家电制造业
	57	68	31	44
	66	39	49	51
	49	29	21	65
	40	45	34	77
	34	56	40	58
	53	51		
	44			
均值	49	48	35	59

- 计算全部观测值的总均值：47.87

单因素方差分析：以各行业服务质量为例说明

在方差分析中，需要计算三个误差平方和，总平方和、组间平方和和组内平方和。

1）总平方和：sum of squares for total，记为SST

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

例如，上述案例中，计算总平方和为：

$$SST = (57 - 47.870)^2 + \cdots + (58 - 47.870)^2 = 4164.609$$

总平方和反映了全部观测值与其均值之间的差异。

说明：计算不要求掌握，Excel可以自动给出这些计算结果。

2）组间平方和：sum of squares for factor A，记为SSA

SSA是各组均值与总均值的误差平方和，反映各样本均值之间的差异程度，因此又称为因素平方和。

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

根据上表中的数据，计算组间平方和为：

$$\begin{aligned} SSA &= \sum_{i=1}^{4} n_i (\bar{x}_i - \bar{\bar{x}})^2 \\ &= 7 \times (49 - 47.870)^2 + 6 \times (48 - 47.870)^2 + 5 \times (35 - 47.870)^2 + 5 \times (59 - 47.870)^2 \\ &1456.609 \end{aligned}$$

3）组内平方和，sum of squares for error，记为SSE。

组内平方和是每个组内的样本数据与其组均值的误差平方和，反映了每个样本各观测值的离散程度。

反映了随机误差的大小。

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

在上面这个例子中，先求出每个行业被投诉的次数与其均值的误差平方和，然后将四个行业的误差平方和加总，即为SSE。

例如，零售业：

$$\sum_{j=1}^{7} (x_{1j} - \bar{x}_1)^2 = (57 - 49)^2 + (66 - 49)^2 + \cdots + (44 - 49)^2 = 700$$

按照相同的方式，可以将其他三个行业的误差平方和都计算出来。

- 旅游业：924
- 航空公司：434
- 家电制造业：650

加总，得

$$SSE = 700 + 924 + 434 + 650 = 2708$$

上面提到的三个平方和之间的关系为：

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

说白了，就是：总平方和=组间平方和+组内平方和。 4164.609=1456.609+2708

接着，将以上计算出来的误差平方和进行平均，即用各平方和除以它们的自由度，将这个结果称为**均方**，其实就是方差。

- SST的自由度为n-1，n为全部观测值的个数。
- SSA的自由度为k-1，k为因素水平的个数。
- SSE的自由度为n-k。

由于需要比较的是组间均方和组内均方之间的差异，所以通常只需要计算SSA的均方和SSE的均方。

首先，来看SSA的均方，SSA的均方也称为组间均方或组间方差，记为MSA。

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

例如，$MSA = \frac{SSA}{k-1} = \frac{1456.609}{4-1} = 485.536$

接着，SSE的均方也称为组内均方或组内方差，记为MSE。

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

例如，$MSE=\frac{SSE}{n-k}=\frac{2708}{23-4}=142.526$

最后，将MSA和MSE进行对比，方差的比值服从F分布。所以，

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

对于上述例子，有

$$F=\frac{MSA}{MSE}=\frac{485.536}{142.526}=3.406$$

统计决策

3、统计决策

取显著性水平$\alpha = 0.05$，根据Excel公式计算出临界值$F_{0.05}(3,19) = 3.13$

由于$F = 3.406 > F_{\alpha}$，所以拒绝原假设，说明不同行业对被投诉次数有显著影响。

4、方差分析表

将以上计算结果整理成一张表，称为方差分析表，一般用Excel做方差分析会自动给出这个表。

差异源	SS	df	MS	F	P-value	F crit
组间	1456.608696	3	485.5362	3.40664269	0.038765	3.127350005
组内	2708	19	142.5263			
总计	4164.608696	22				

利用P值决策，如果P值小于$\alpha = 0.05$，则拒绝原假设。

单因素方差分析

主要内容：

- 单因素方差分析的数据结构
- 单因素方差分析3步法
- 单因素方差分析效果评估
- 用Excel做单因素方差分析
- 方差分析中的多重比较：LSD检验法

评估方差分析的效果

根据前面的分析，可以结论：不同行业跟被投诉次数是有关系的。

如何来度量这种关系的强度呢？可以用组间平方和占总平方和的比例大小来反映，即

$$R^2 = \frac{SSA}{SST}$$

例如，对于上述案例，有

$$R^2 = \frac{SSA}{SST} = \frac{1456.6087}{4164.6087} = 0.341$$

结论：

说明行业对被投诉次数的影响占总效应的34.1%，或者，行业对被投诉次数差异解释的比例达到近35%。

单因素方差分析

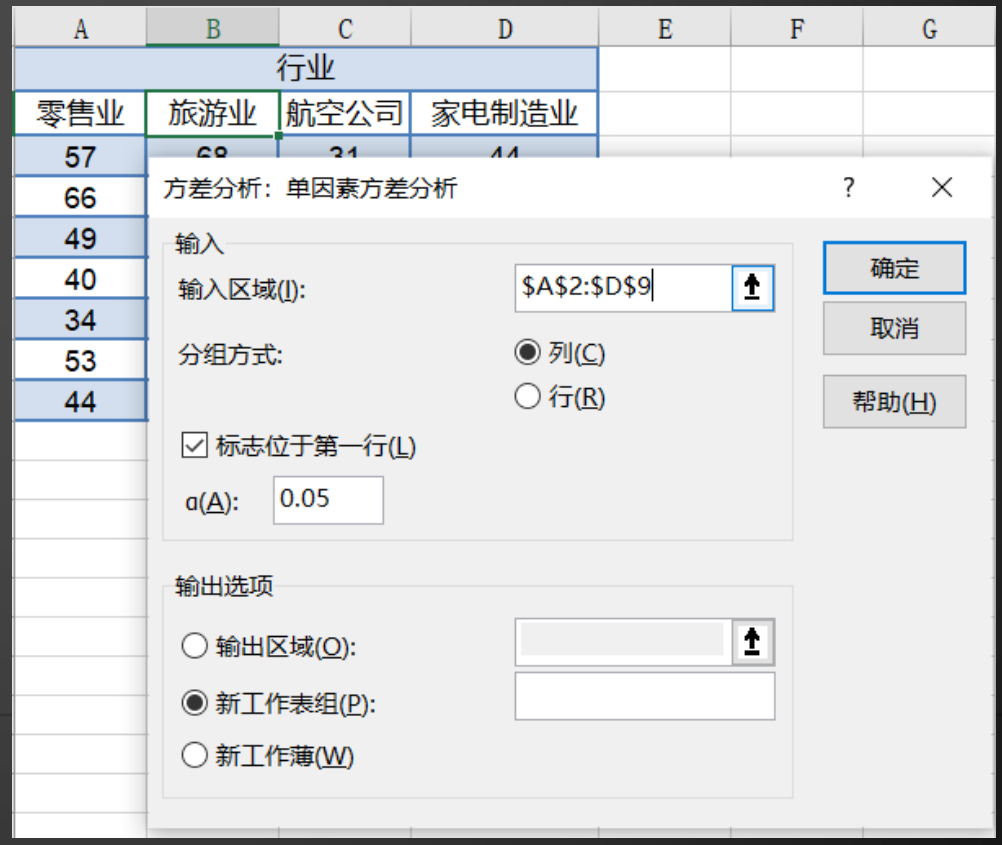
主要内容：

- 单因素方差分析的数据结构
- 单因素方差分析3步法
- 单因素方差分析效果评估
- 用Excel做单因素方差分析
- 方差分析中的多重比较：LSD检验法

用Excel进行方差分析

前面的方差分析中，是通过手动计算来得出结论的。实际工作中，可以用工具进行方差分析。步骤如下。

1. 菜单【数据】-【数据分析】，在弹出的“分析工具”对话框中，选择【单因素方差分析】
2. 在弹出的对话框中，做如下选择：
   - 【输入区域】选择单元格区域A2:D9
   - 勾选【标志位于第一行】
   - 【输出选项】用默认的“新工作表组”



用Excel进行方差分析

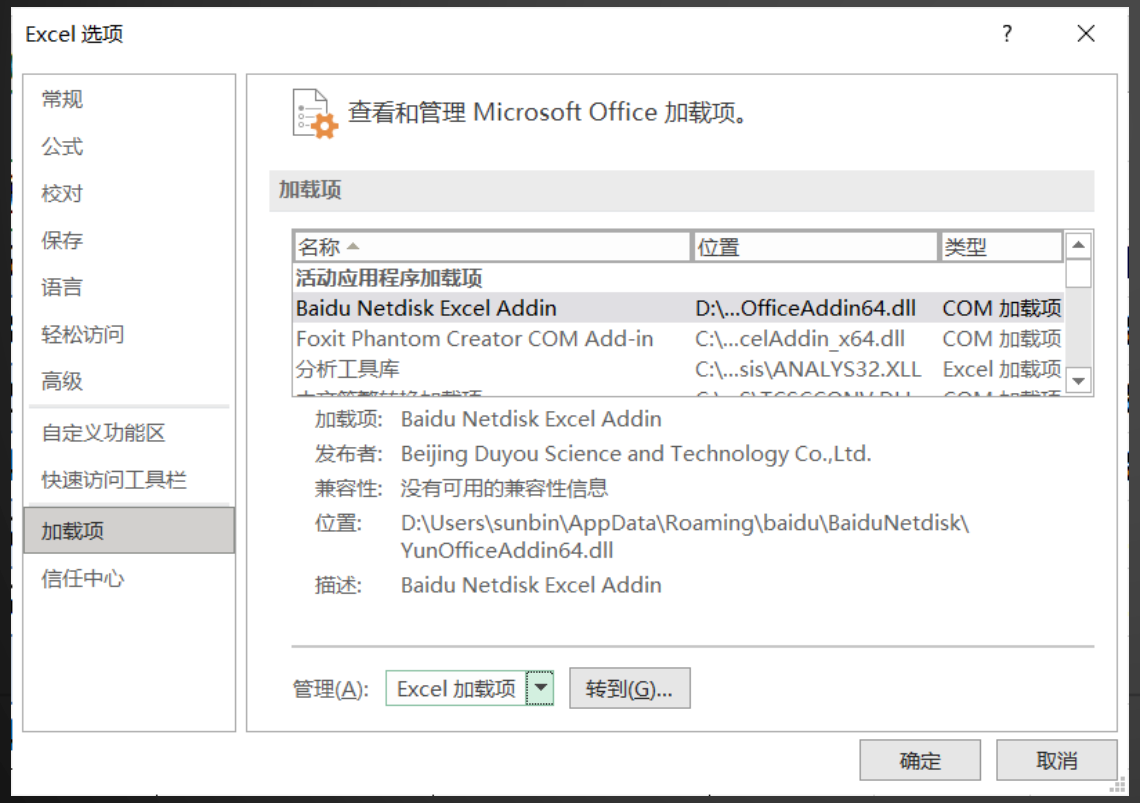
- 单击【确定】按钮后，得到如下输出结果。

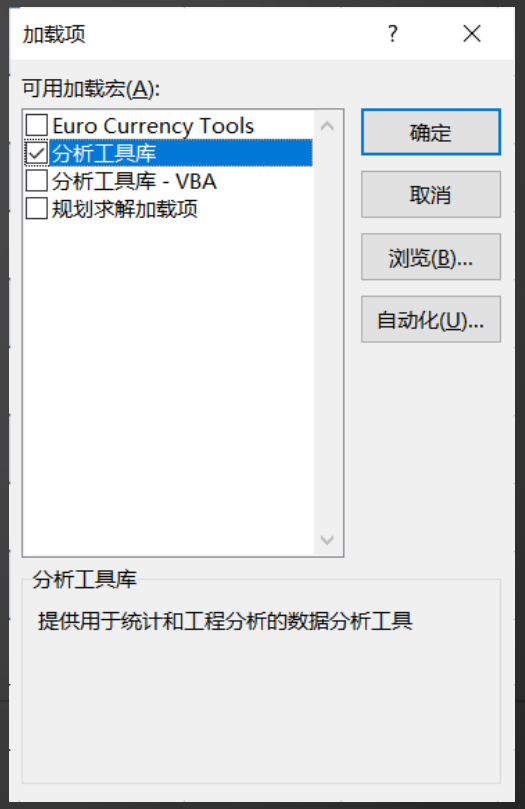
A	B	C	D	E	F	G
方差分析：单因素方差分析						
SUMMARY						
组	观测数	求和	平均	方差		
零售业	7	343	49	116.6667		
旅游业	6	288	48	184.8		
航空公司	5	175	35	108.5		
家电制造业	5	295	59	162.5		
方差分析						
差异源	SS	df	MS	F	P-value	F crit
组间	1456.609	3	485.5362	3.406643	0.038765	3.12735
组内	2708	19	142.5263			
总计	4164.609	22				

说明：一般是通过P值进行判断，如果P值小于$\alpha = 0.05$，则拒绝原假设

附：如何调出Excel中的数据分析工具库

- 菜单：【文件】-【选项】
- 在弹出的【Excel选项】对话框中，选择“加载项”，选择“Excel加载项”
- 在弹出的【加载项】对话框中，勾选“分析工具库”，单击【确定】按钮即可。





单因素方差分析

主要内容：

- 单因素方差分析的数据结构
- 单因素方差分析3步法
- 单因素方差分析效果评估
- 用Excel做单因素方差分析
- 方差分析中的多重比较：LSD检验法

方差分析中的多重比较：LSD检验法

- 采用的方法就是多重比较法，即通过对总体均值之间的配对比较来进一步检验到底哪些行业之间存在差异。多重比较法有很多种方法，最常用的一种是LSD法，least significant difference，最小显著差异方法，LSD法对行业进行两两检验。

方差分析中的多重比较：LSD检验法

LSD检验方法的步骤如下。

1. 提出假设：$H_0：\mu_i = \mu_j$；$H_1：\mu_i \neq \mu_j$
2. 计算检验统计量：$\overline{x}_i - \overline{x}_j$
3. 计算LSD，公式为：

$$LSD = t_{\alpha/2}(n-k)\sqrt{MSE(\frac{1}{n_i}+\frac{1}{n_j})}$$

其中，$t_{\alpha/2}$为t分布的临界值，自由度为n-k；MSE为组内方差。

4. 决策：如果$|\overline{x}_i - \overline{x}_j| > LSD$，则拒绝$H_0$；如果$|\overline{x}_i - \overline{x}_j| < LSD$，则接受$H_0$。

方差分析中的多重比较：LSD检验法

【例】接下来以前面行业被投诉次数为例来说明LSD检验法。

第一步，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：$H_0: \mu_3 = \mu_4$；　$H_1: \mu_3 \neq \mu_4$

零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			

方差分析中的多重比较：LSD检验法

第二步，计算检验统计量。

- $|\overline{x}_1 - \overline{x}_2| = |49 - 48| = 1$
- $|\overline{x}_1 - \overline{x}_3| = |49 - 35| = 14$
- $|\overline{x}_1 - \overline{x}_4| = |49 - 59| = 10$
- $|\overline{x}_2 - \overline{x}_3| = |48 - 35| = 13$
- $|\overline{x}_2 - \overline{x}_4| = |48 - 59| = 11$
- $|\overline{x}_3 - \overline{x}_4| = |35 - 59| = 24$

零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			
49	48	35	59

方差分析中的多重比较：LSD检验法

第三步，计算LSD。之前已经计算出，$MSE = 142.526$。根据样本自由度$n - k = 23 - 4 = 19$，$t_{\alpha/2} = 2.093$，所以，各检验的LSD如下。

- 检验1：$LSD_1 = 2.903 \times \sqrt{142.526 \times (\frac{1}{7} + \frac{1}{6})} = 13.90$
- 检验2：$LSD_2 = 2.903 \times \sqrt{142.526 \times (\frac{1}{7} + \frac{1}{5})} = 14.63$
- 检验3：$LSD_3 = LSD_2 = 14.63$
- 检验4：$LSD_4 = 2.903 \times \sqrt{142.526 \times (\frac{1}{6} + \frac{1}{5})} = 15.13$
- 检验5：$LSD_5 = LSD_4 = 15.13$
- 检验6：$LSD_6 = 2.903 \times \sqrt{142.526 \times (\frac{1}{5} + \frac{1}{5})} = 15.80$

A	B	C	D
方差分析：单因素方差分析			
SUMMARY			
组	观测数	求和	平均
零售业	7	343	49
旅游业	6	288	48
航空公司	5	175	35
家电制造业	5	295	59
方差分析			
差异源	SS	df	MS
组间	1456.609	3	485.5362
组内	2708	19	142.5263
总计	4164.609	22	

方差分析中的多重比较：LSD检验法

第四步：做出决策。

- $|\overline{x}_1 - \overline{x}_2| = 1 < 13.90$，不能拒绝$H_0$，即零售业与旅游业被投诉次数之间没有显著差异。
- $|\overline{x}_1 - \overline{x}_3| = 14 < 14.63$，不能拒绝$H_0$，即零售业与航空公司被投诉次数之间没有显著差异。
- $|\overline{x}_1 - \overline{x}_4| = 10 < 14.63$，不能拒绝$H_0$，即零售业与家电制造业被投诉次数之间没有显著差异。
- $|\overline{x}_2 - \overline{x}_3| = 13 < 15.13$，不能拒绝$H_0$，即旅游业与航空公司被投诉次数之间没有显著差异。
- $|\overline{x}_2 - \overline{x}_4| = 11 < 15.13$，不能拒绝$H_0$，即旅游业与家电制造业被投诉次数之间没有显著差异。
- $|\overline{x}_3 - \overline{x}_4| = 24 > 15.80$，拒绝$H_0$，即航空公司与家电制造业被投诉次数之间有显著差异。

本章主要内容

- 方差分析基本原理
- 单因素方差分析
- 双因素方差分析

双因素方差分析

主要内容：

- 双因素方差分析及其数据结构
- 双因素方差分析中的两类问题
- 双因素方差分析3步法
- 用Excel做双因素方差分析

双因素方差分析

主要内容：

- 双因素方差分析及其数据结构
- 双因素方差分析中的两类问题
- 双因素方差分析3步法
- 用Excel做双因素方差分析

双因素方差分析

- 当方差分析中涉及两个分类型自变量时，称为双因素方差分析（two-way analysis of variance）
- 【例】有4个品牌的彩电在5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。

		地区因素				
		地区1	地区2	地区3	地区4	地区5
品牌因素	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

行因素 A	列因素 B			
	B_1	B_2	…	B_k
A_1	x_{11}	x_{12}	…	x_{1k}
A_2	x_{21}	x_{22}	…	x_{2k}
…	…	…	…	…
A_r	x_{r1}	x_{r2}	…	x_{rk}

双因素方差分析的数据结构

双因素方差分析

主要内容：

- 双因素方差分析及其数据结构
- 双因素方差分析中的两类问题
- 双因素方差分析3步法
- 用Excel做双因素方差分析

双因素方差分析中的两类问题

在双因素方差分析中：

- 如果两个因素相互独立，此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。
- 如果两个因素结合后产生新效应，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，或称为可重复双因素方差分析。

		地区因素				
		地区1	地区2	地区3	地区4	地区5
品牌因素	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

无交互作用的双因素方差分析

		路段（列变量）	
		路段1	路段2
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

有交互作用的双因素方差分析

双因素方差分析

主要内容：

- 双因素方差分析及其数据结构
- 双因素方差分析中的两类问题
- 双因素方差分析3步法
- 用Excel做双因素方差分析

双因素方差分析3步法

1、提出假设

由于双因素方差分析涉及两个因素，所以需要分别对两个因素提出假设。

- 行因素的假设
- 列因素的假设

2、构造检验的统计量

针对行因素和列因素，分别计算误差平方和——>方差——>方差比——>F分布

计算出F值

3、统计决策

根据显著性水平α计算出临界值，和第2步计算出来的F值比较

根据比较结果，作出决策，接受还是拒绝原假设

双因素方差分析基本步骤

1、提出假设

由于双因素方差分析涉及两个因素，所以需要分别对两个因素提出假设：

对行因素提出假设：

$H_0：\mu_1 = \mu_2 = \cdots = \mu_k$ 行因素对因变量没有显著影响

$H_1：\mu_1, \mu_2, \cdots, \mu_k$不全相等 行因素对因变量有显著影响

对列因素提出假设：

$H_0：\mu_1 = \mu_2 = \cdots = \mu_k$ 列因素对因变量没有显著影响

$H_1：\mu_1, \mu_2, \cdots, \mu_k$不全相等 列因素对因变量有显著影响

双因素方差分析基本步骤

2、构造检验的统计量

从总平方和的分解入手。

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

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

其中，分解后的等式右边第一项为行因素产生的误差平方和，记为SSR，即

$$SSR=\sum_{i=1}^{k}\sum_{j=1}^{r}(x_{i.}-\bar{\bar{x}})^2$$

双因素方差分析基本步骤

第二项是列因素产生的误差平方和，记为SSC，即

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

第三项是随机误差平方和，记为SSE，即

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

上述平方和的关系为：

$$SST = SSR + SSC + SSE$$

接着，可以计算出下列各均方：

行因素的均方，记为MSR，即

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

列因素的均方，记为MSC，即

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

随机误差项的均方，记为MSE，即

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

为了检验行因素和列因素对因变量的影响是否显著，采用F统计量：

行因素：$F_R = \frac{MSR}{MSE} \sim F(k-1, (k-1)(r-1))$

列因素：$F_C = \frac{MSC}{MSE} \sim F(r-1, (k-1)(r-1))$

3、统计决策

给定显著性水平α，将上一步就算出来的F值与临界值进行比较。

- 行因素：若$F_R > F_\alpha$，则拒绝原假设，反之，接受原假设。
- 列因素：若$F_C > F_\alpha$，则拒绝原假设，反之，接受原假设。

最后，可以将以上分析结果列成一张表，即双因素方差分析表。

差异源	SS	df	MS	F	P-value	F crit
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				

说明：实际应用，用Excel做双因素方差分析，自动给出上述结果。

双因素方差分析

主要内容：

- 双因素方差分析及其数据结构
- 双因素方差分析中的两类问题
- 双因素方差分析3步法
- 用Excel做双因素方差分析

用Excel进行双因素方差分析

		地区因素				
		地区1	地区2	地区3	地区4	地区5
品牌因素	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

1、对两个因素分别提出如下假设

行因素：品牌

$H_0：\mu_1 = \mu_2 = \mu_3 = \mu_4$　　品牌对销售量没有显著影响

$H_1：\mu_1, \mu_2, \mu_3, \mu_4$不全相等　　品牌对销售量有显著影响

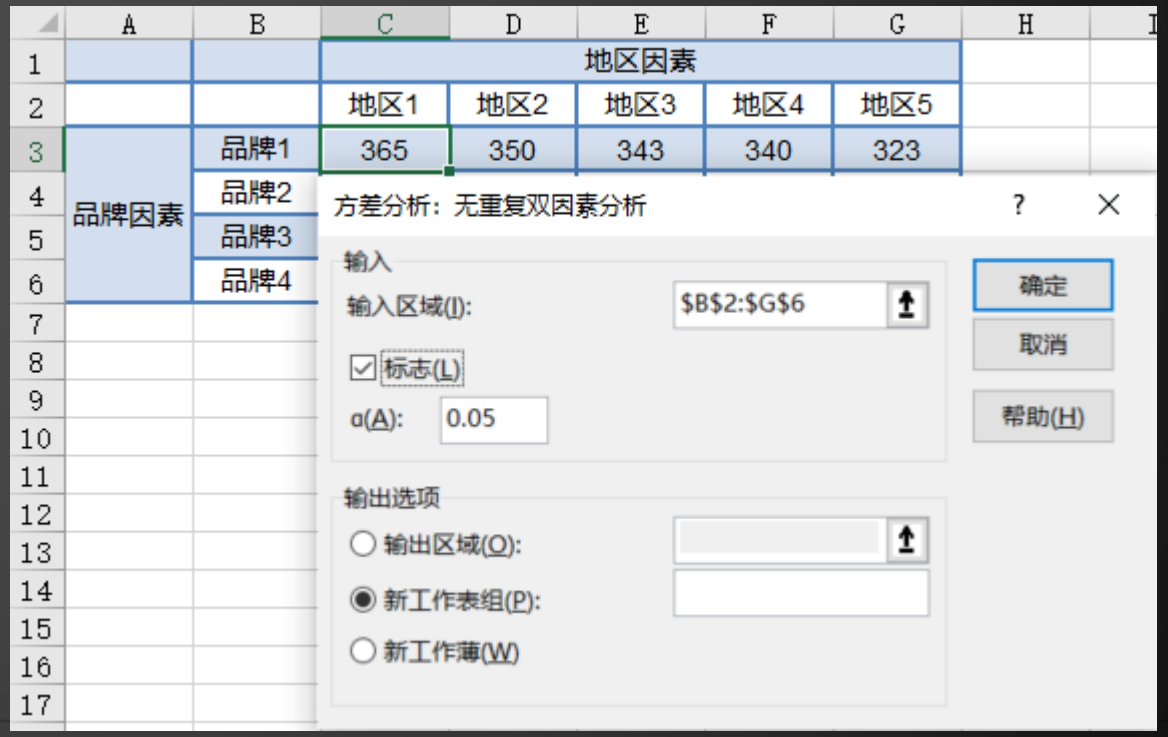
列因素：地区

$H_0：\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$　　地区对销售量没有显著影响

$H_1：\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$不全相等　　地区对销售量有显著影响

用Excel进行双因素方差分析

1. 菜单【数据】-【数据分析】，在弹出的“分析工具”对话框中选择【无重复双因素方差分析】
2. 在弹出的对话框中，做如下选择：
   - 【输入区域】选择单元格区域B2:G6
   - 勾选【标志】
   - 【输出选项】用默认的“新工作表组”



用Excel进行双因素方差分析

- 结论：

方差分析：无重复双因素分析

SUMMARY	观测数	求和	平均	方差
品牌1	5	1721	344.2	233.7
品牌2	5	1739	347.8	295.7
品牌3	5	1685	337	442.5
品牌4	5	1424	284.8	249.2
地区1	4	1356	339	1224.667
地区2	4	1321	330.25	1464.25
地区3	4	1357	339.25	822.9167
地区4	4	1273	318.25	1538.917
地区5	4	1262	315.5	241.6667

方差分析

差异源	SS	df	MS	F	P-value	F crit
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				

说明：一般是通过P值进行判断，如果P值小于$\alpha = 0.05$，则拒绝原假设

有交互作用的双因素方差分析

【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，让一个名交通警察分别在两个路段的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：分钟）的数据，如下表所示。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。$(\alpha = 0.05)$

		路段（列变量）	
		路段1	路段2
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

1、提出假设：行因素、列因素、交互作用因素

行因素：时段

$H_0：\mu_1 = \mu_2$　　时段对行车时间没有显著影响

$H_1：\mu_1 \neq \mu_2$　　时段对行车时间有显著影响

列因素：路段

$H_0：\mu_1 = \mu_2$　　路段对行车时间没有显著影响

$H_1：\mu_1 \neq \mu_2$　　路段对行车时间有显著影响

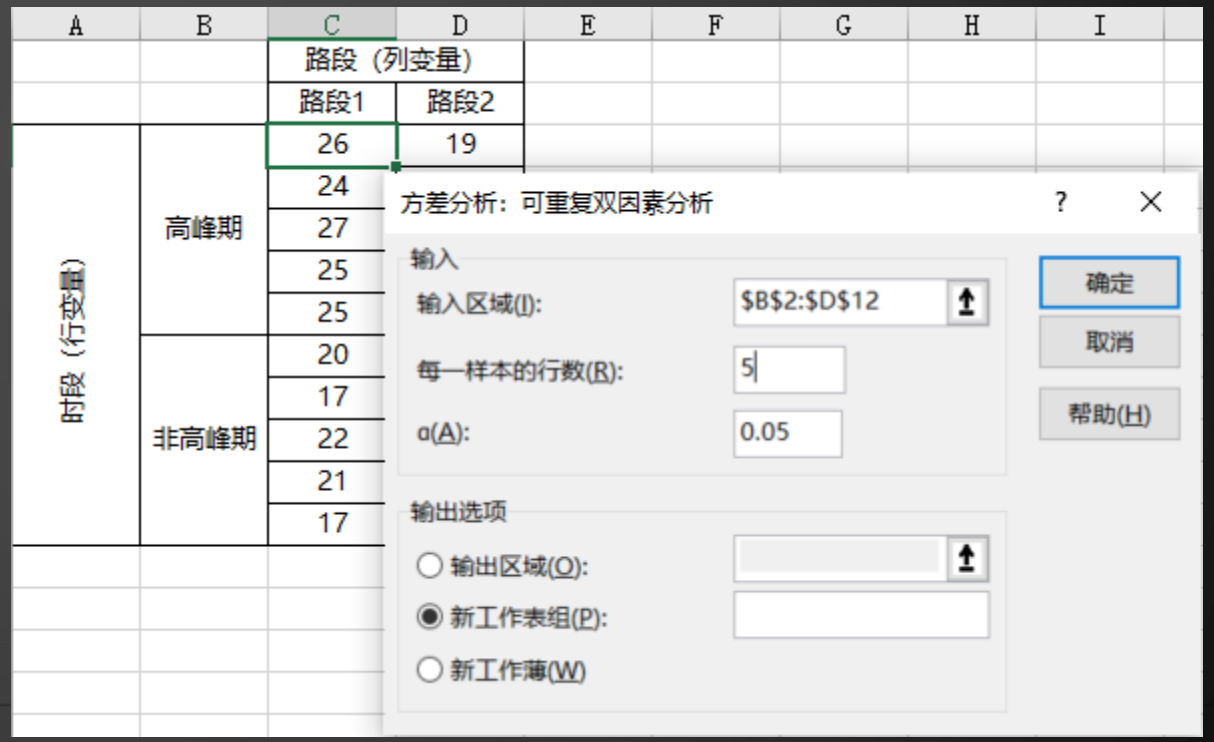
交互作用因素：

$H_0：\mu_{11} = \mu_{12} = \mu_{21} = \mu_{22}$　　交互作用对行车时间没有显著影响

$H_1：\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}$不全相等　　交互作用对行车时间有显著影响

用Excel进行双因素方差分析

1. 菜单【数据】-【数据分析】，在弹出的“分析工具”对话框中选择【可重复双因素方差分析】
2. 在弹出的对话框中，做如下选择：
   - 【输入区域】选择单元格区域B2:G6
   - 【每一样本的行数】
   - 【输出选项】用默认的“新工作表组”



用Excel进行双因素方差分析

方差分析：可重复双因素分析

SUMMARY	路段1	路段2	总计
高峰期			
观测数	5	5	10
求和	127	105	232
平均	25.4	21	23.2
方差	1.3	2.5	7.066667
非高峰期			
观测数	5	5	10
求和	97	76	173
平均	19.4	15.2	17.3
方差	5.3	6.7	10.23333
总计			
观测数	10	10	
求和	224	181	
平均	22.4	18.1	
方差	12.93333	13.43333	

方差分析

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				

说明：一般是通过P值进行判断，如果P值小于$\alpha = 0.05$，则拒绝原假设