第 7 章　方差分析

内容提要

7.1 方差分析引论

- 什么是方差分析
- 有关术语
- 基本思想与原理
- 基本假定
- 假设提出

什么是方差分析

定义

方差分析(analysis of variance,ANOVA),是指通过检验各个总体的均值是否相等，来判断**分类型变量**与**数值型变量**是否存在相关性。

起源： 20世纪20年代，英国统计学家费希尔进行实验设计时，为了解释实验数据而引入。

应用： 心理学、生物学、工程、医药…

形式： 通过分析数据的误差，比较多个总体的均值是否相等

实质： 分析**分类型变量**与**数值型变量**之间的相关性。

分类

- 单因素方差分析：**一个分类变量**与一个数值变量
- 双因素方差分析：**两个分类变量**与一个数值变量
  - **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

方差分析：例子

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：

表 10-1　　四个行业被投诉次数

	B	C	D	E
1	行业			
2	零售业	旅游业	航空公司	家电制造业
3	57	68	31	44
4	66	39	49	51
5	49	29	21	65
6	40	45	34	77
7	34	56	40	58
8	53	51		
9	44			

有关术语

因素（Factor）

也称因子，是指所要检验的对象（**分类变量**）。

例子：行业

水平（Treatment）

也称处理，是指因素的不同表现（**分类变量的各个类型**）。

例子：行业的类型包括零售业、旅游业、航空业、家电制造业

观察值

每个因素水平下的样本数据（**数值型数据的取值**）。

例子：零售业的投诉次数，旅游业的投诉次数…

基本思想与原理

均值比较

- 比较多个总体的均值是否相等。
- 例子：4个行业的投诉均值是否相等

均值比较的含义

- 分析**分类型变量**与**数值型变量**之间的**关系**。
- 例子：行业与投诉次数是否有关

分析方法

- 散点图：初步判断
- 方差分析：精确的检验

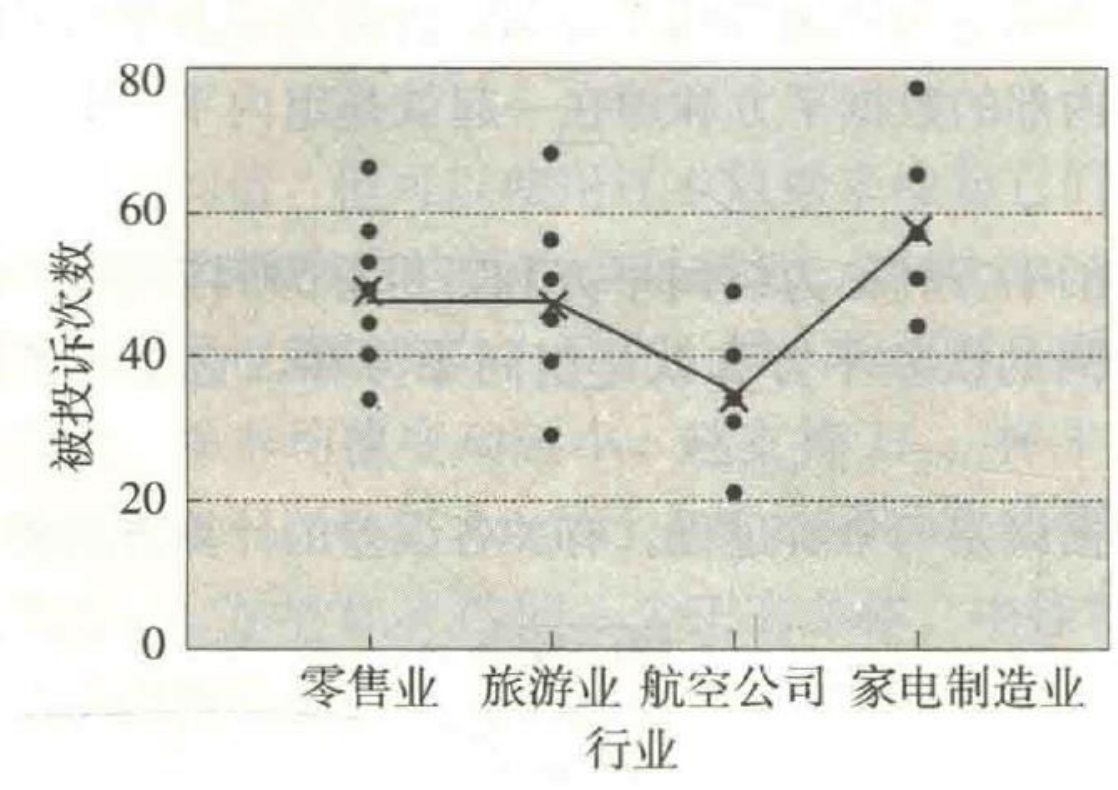

图 10－1　不同行业被投诉次数的散点图

方差分析：误差类型

两种类型的误差

- **随机误差**
  - 含义：因素的某一水平下的，样本观测值之间的差异。
  - **决定：抽样随机性**
  - 例子：零售业各个企业的投诉次数的差异主要与抽样随机性有关
- **系统误差**
  - 含义：因素的不同水平之间的观测值差异。
  - 决定：**因素水平的差异（系统差异）**
  - 例子：零售业与旅游业的投诉次数差异，可能与**抽样随机性**有关，更有可能与**行业差异**有关。

方差分析：误差分解与误差分析

误差分解　　SST=SSA+SSE

- **总误差（SST）**
  - 包括：全部误差（随机误差和系统误差）
  - 例子：全部企业的误差平方和
- **组内误差（SSE）**
  - 仅包括：**随机误差**
  - 例子：零售业投诉次数的误差平方和
- **组间误差（SSA）**
  - 包括：**随机误差**和**系统误差**
  - 例子：4个行业投诉次数之间的误差平方和

误差分析

- 若行业与投诉次数**没有关系**，则组间误差（SSA）只包含**随机误差**，没有系统误差，此时，组间误差（SSA）的均值和组内误差（SSE）的均值之比（F统计量），**接近1**。
- 若行业与投诉次数**有关系**，则组间误差（SSA）同时包含**随机误差**和**系统误差**，此时，组件误差（SSA）的均值和组内误差（SSE）的均值之比（F统计量），**大于1**。
- 当上述比值足够大时（**F统计量大于临界值**），可以判断：**行业与投诉次数有关**。

基本假定

- **每个总体都应服从正态分布**

  例：每个行业的投诉次数都服从正态分布

- **各个总体的方差必须相同**

  例：4个行业的投诉次数方差相同

- **观测值是独立的**

  例：每个行业的投诉次数独立于其他行业的投诉次数

假设的提出

假设的一般提法：

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，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：$\mu_1=\mu_2=\cdots=\mu_k$　　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

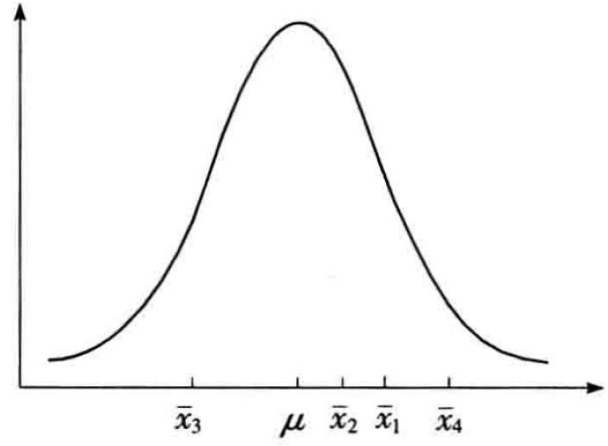

图 10—3　H_0 为真时 x 的抽样分布

例子：

在例 10.1 中，设零售业被投诉次数的均值为 μ_1，旅游业被投诉次数的均值为 μ_2，航空公司被投诉次数的均值为 μ_3，家电制造业被投诉次数的均值为 μ_4。为检验行业对被投诉次数是否有影响，需要提出如下假设：

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：μ_1，μ_2，μ_3，μ_4 不全相等　　行业对被投诉次数有显著影响

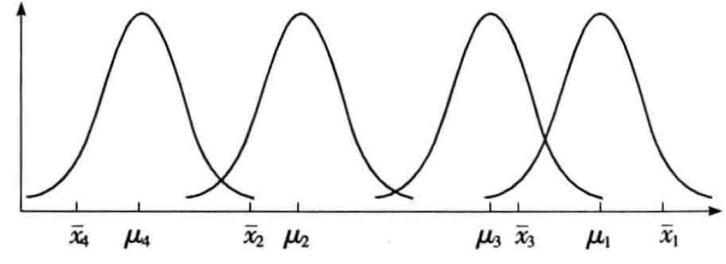

图 10—4　4 个行业被投诉次数的均值全不相同时 x 的抽样分布

7.2 单因素方差分析

数据结构

表 10—2　　单因素方差分析的数据结构

	A	B	C	D	E
1	观测值	因素(i)			
2	(j)	A_1	A_2	…	A_k
3	1	x_{11}	x_{21}	…	x_{k1}
4	2	x_{12}	x_{22}	…	x_{k2}
5	⋮	⋮	⋮	⋮	⋮
6	n	x_{1n}	x_{2n}	…	x_{kn}

方差分析表

表 10—4　　方差分析表的一般形式

	A	B	C	D	E	F	G
1 2	误差来源	平方和 SS	自由度 df	均方 MS	F值	P值	F临界值
3	組间 (因素影响)	SSA	k-1	MSA	MSA/MSE		
4	組内 (误差)	SSE	n-k	MSE			
5	总和	SST	n-1				

分析步骤

- **提出假设**

H_0：$\mu_1=\mu_2=\cdots=\mu_i=\cdots=\mu_k$　　自变量对因变量没有显著影响

H_1：$\mu_i(i=1, 2, \cdots, k)$不全相等　　自变量对因变量有显著影响

- **构造检验统计量**
  - 样本均值
  - 总均值
  - 误差平方和：SST SSA SSE
  - 计算统计量
  - 形成**方差分析表**
- **做出统计决策**
  - F统计量**大于**临界值，**拒绝**H_0
  - F统计量**小于**临界值，**接受**H_0

单因素方差分析：例子

数据结构

表 10—3　　四个行业被投诉次数及其均值

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

分析步骤

- 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：μ_1，μ_2，μ_3，μ_4 不全相等　　行业对被投诉次数有显著影响

- 构造检验统计量

表 10—5　　四个行业被投诉次数的方差分析表

	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	1456.6087	3	485.5362	3.4066	0.0388	3.1274
4	组内	2708	19	142.5263			
5							
6	总计	4164.6087	22				

- 做出统计决策

F统计量**大于**临界值，**拒绝**H_0

即行业对被投诉次数有显著影响。

单因素方差分析：关系强度与多重比较

关系强度的测量

- 方差分析的不足
  - 只能判断：**是否有关系**
  - **无法判断：关系的强度**
- 关系强度的测量

$$R^2=\frac{SSA(\text{组间 }SS)}{SST(\text{总 }SS)}$$

- 例子

$$R^2=\frac{SSA(\text{组间 }SS)}{SST(\text{总 }SS)}=\frac{1\ 456.608\ 696}{4\ 164.608\ 696}=0.349\ 759=34.975\ 9\%$$

多重比较

- 方差分析的不足
  - 只能判断：均值是否相等
  - **无法判断：哪两个均值不相等**
- 多重比较的原理

  通过配对比较，具体判断哪两个均值之间存在不相等。

7.3 双因素方差分析

无交互作用的双因素方差分析

表 10—9　　双因素方差分析表

	A	B	C	D	E	F	G
1	误差来源	误差平方和 SS	自由度 df	均方 MS	F 值	P 值	F临界值
2	行因素	SSR	k-1	MSR	F_R		
3	列因素	SSC	r-1	MSC	F_C		
4	误差	SSE	(k-1)×(r-1)	MSE			
5	总和	SST	kr-1				

无交互作用的双因素方差分析

也称**无重复双因素分析**，是指两个分类变量对数值变量的影响是**相互独立**的。

有交互作用的双因素方差分析

表 10—14　　有交互作用的双因素方差分析表的结构

	A	B	C	D	E	F	G
1	误差来源	平方和 SS	自由度 df	均方 MS	F 值	P 值	F 临界值
2	行因素	SSR	k-1	$MSR = \frac{SSR}{k-1}$	$F_R = \frac{MSR}{MSE}$		
3	列因素	SSC	r-1	$MSC = \frac{SSC}{r-1}$	$F_C = \frac{MSC}{MSE}$		
4	交互作用	SSRC	(k-1)(r-1)	$MSRC = \frac{SSRC}{(k-1)(r-1)}$	$F_{RC} = \frac{MSRC}{MSE}$		
5	误差	SSE	kr(m-1)	$MSE = \frac{SSE}{kr(m-1)}$			
6	总和	SST	n-1				

有交互作用的双因素方差分析

也称**可重复双因素分析**，是指两个分类变量对数值变量的影响是**不独立**的，存在**交互作用**。

本章小结