



北京航空航天大学

— 经济管理学院 —

BEIHANG UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

第四讲：方差分析

康雁飞

数量经济与商务统计系

Outline

- 1 一个例子
- 2 方差分析的本质及误差分解
- 3 单因素方差分析
- 4 双因素方差分析 (拓展内容)

Outline

- 1 一个例子
- 2 方差分析的本质及误差分解
- 3 单因素方差分析
- 4 双因素方差分析 (拓展内容)

案例：行业服务质量是否有显著差异

消费者对 23 家企业投诉次数如下。

消费者对四个行业的投诉次数				
	行业			
观测值	零售业	旅游业	航空公司	家电制造业
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			



问题

1 这四个行业之间的服务质量是否有显著差异？

2 即判断“行业”对“被投诉次数”是否有显著影响？

“行业”是分类变量，“被投诉次数”是数值变量；问题转化为：分类变量对数值变量的影响？

问题与回顾

- 1 如何衡量分类变量对数值变量的影响？
- 2 行业对被投诉次数是否有显著影响？
- 3 来自不同地区的大学生每个月的平均生活费支出是否相同？
- 4 家电的品牌对它们的销售量是否有影响？

前面几章讨论的都是一个总体或两个总体的统计分析问题, 如何进行多个总体的统计分析, 尤其是多总体均值的比较问题？

多总体均值比较能否使用 t 检验？

- 1 需要进行多次比较： t 检验一次研究两个样本；案例中 4 个总体，意味着要进行 6 次两样本的 t 检验
- 2 增加犯第一类错误的概率，假设每次 t 检验犯第一类错误的概率是 0.05，那么整体犯第一类错误的概率将达到：
$$1 - (1 - 0.05)^6 = 0.265$$
- 3 警惕过多的检验

Outline

- 1 一个例子
- 2 方差分析的本质及误差分解
- 3 单因素方差分析
- 4 双因素方差分析 (拓展内容)

方差分析

- 方差分析 (Analysis of Variance, ANOVA): 20 世纪 20 年代英国统计学家费希尔在进行试验设计时为解释试验数据首次提出
- 表面理解: 方差分析是检验多个总体均值是否相等的统计方法:

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

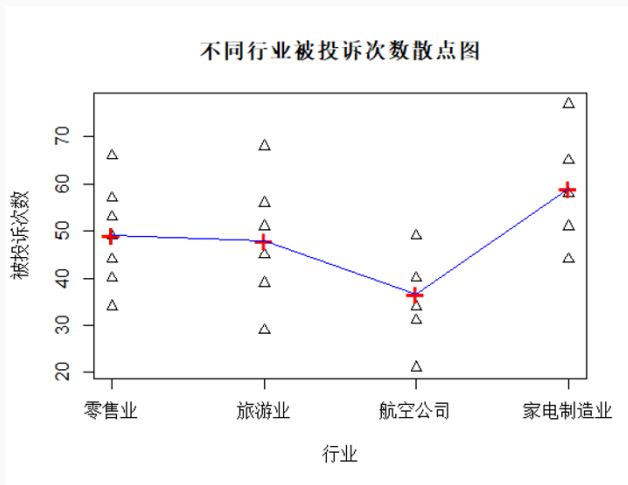
方差分析

- 本质：方差分析研究分类变量对数值变量的影响
- 含义：通过检验总体的均值是否相等来判断分类变量对数值变量是否有显著影响
 - ▶ 分析 4 个行业之间的服务质量是否存在显著差异，也就是要判断“行业”对“被投诉次数”是否有显著影响
 - ▶ 作出这种判断归结为检验这四个行业被投诉次数的总体均值是否相等
 - ▶ 若它们的均值相等，则意味着“行业”对被投诉次数是没有影响的，即它们之间的服务质量没有显著差异；若均值不全相等，则意味着“行业”对被投诉次数是有影响的，它们之间的服务质量有显著差异

方差分析：相关术语

- 因素或因子 (factor): 所要检验的对象. 例如 “行业”
- 水平 (level) 或处理 (treatment): 因子的不同表现. 例如: 零售业、旅游业、航空公司、家电制造业 (单因素 4 水平试验)
- 观测值 (Observation): 在每个因素水平下得到的样本数据

方差分析：图形描述



方差分析：误差分解

- 方差：度量数据波动性或误差大小（信息量）
- 判断不同总体的均值是否相等需要分析：
 - 1 差异是由抽样的随机性导致的？
 - 2 差异是由数据来自不同总体造成的？
- 误差分解
 - 1 总误差
 - 2 组内误差
 - 3 组间误差

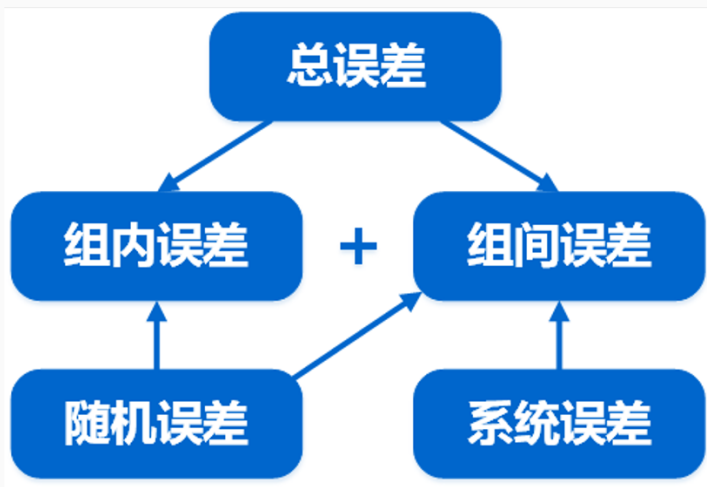
- 1 总误差 (Total error): 反映全部观测数据的误差 (案例: 所抽取的 23 家企业被投诉次数之间的误差)

- 2 组内误差 (Within-class error): 因素的同一水平下, 样本各观察值之间的差异 (同一行业下 (比如: 零售业) 不同企业被投诉次数之间的差异, 反映一个样本内部数据的离散程度)
- ▶ 企业随机抽取, 它们之间的差异可以看出是随机因素的影响造成的, 也称为随机误差 (Random error)

3 组间误差 (Between-class error): 因素的不同水平之间观察值的差异, 也称为处理误差

- ▶ 不同行业之间 (比如: 航空公司和家电制造业) 的被投诉次数之间的差异, 反映不同样本之间数据的离散程度
- ▶ 差异可能是由于抽样的随机性所造成的 (随机误差), 也可能是由于行业本身所造成的 (系统误差), 组间误差是随机误差和系统误差的总和

方差分析：误差分解



如何表示误差?数据的误差用离差平方和表示。

方差分析：误差分解

■ 总误差 ← 总平方和 (SS for total, SST)

- ▶ 反映全部数据误差大小的平方和
- ▶ 案例：所抽取的全部 23 家企业被投诉次数之间的误差平方和，反映了全部观测值的离散状况

$$SST = \sum_{i=1}^4 \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = 4164.609$$

- ▶ x_{ij} 第 i 个行业第 j 家企业被投诉次数; $\bar{x}$ 是全部数据的总均值

方差分析：误差分解

■ 组内误差 ← 误差平方和 (SS of error, SSE)

- ▶ 反映每个样本内部误差大小的平方和
- ▶ 案例：同一行业所抽取企业被投诉次数之间的误差平方和，反映了每个行业观测值的离散状况

$$SSE = \sum_{i=1}^4 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = 2708$$

- ▶ $\bar{x}_i$ 为第 i 个行业被投诉次数的均值
- ▶ 组内平方和 (within-group sum of squares)

方差分析：误差分解

■ 组间误差 ← 处理平方和 (SS of treatment, SSA)

- ▶ 反映样本间误差大小的平方和
- ▶ 案例：4 个行业被投诉次数之间的误差平方和，反映了样本均值之间的差异程度

$$\begin{aligned} SSA &= \sum_{i=1}^4 \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^4 n_i (\bar{x}_i - \bar{x})^2 \\ &= 1456.609 \end{aligned}$$

- ▶ 组间平方和 (between-group sum of squares)

方差分析：误差平方和分解

$$SST = SSE + SSA$$

$$\sum_{i=1}^4 \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^4 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^4 n_i (\bar{x}_i - \bar{x})^2$$

误差分析

- 方差分析就是要分析数据的总误差中有没有处理 (组间) 误差
- 案例：检验被投诉次数的差异主要是由什么原因引起的
- 如果处理 (行业) 对观测数据 (被投诉次数) 没有显著影响, 则意味着没有处理误差, 也就是说每种处理 (行业) 所对应的总体均值应该相等

Outline

- 1 一个例子
- 2 方差分析的本质及误差分解
- 3 单因素方差分析
- 4 双因素方差分析 (拓展内容)

单因素方差分析

- 根据所分析的分类自变量的多少，方差分析可分为单因素方差分析和双因素方差分析
- 开篇案例：只涉及到一个分类变量（行业）和一个数值变量（被投诉次数）
- 单因素方差分析（one-way analysis of variance）：研究一个分类变量（因素）对数值变量（观察指标）的影响

单因素方差分析

处理、水平 (i)				
观	A_1	A_2	$\dots$	A_k
测	x_{11}	x_{21}	$\dots$	x_{k1}
值	x_{12}	x_{22}	$\dots$	x_{k2}
(j)	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_{1n_1}	x_{2n_k}	$\dots$	x_{kn_k}
样本量	n_1	n_2	$\dots$	n_k

单因素方差分析

- 数据结构: 假设因素 A 有 k 个水平 (处理): $A_1, A_2, \dots, A_k$;
- 处理 i 的第 j 个观测值 $x_{ij} (j = 1, 2, \dots, n_i; i = 1, 2, \dots, k)$
- n_i : 处理 i 中所抽取的样本量

单因素方差分析

- 统计模型及假设检验
- 一般步骤
- 方差分析中的多重比较
- 用 R 进行方差分析

单因素方差分析: 统计模型及假设检验

单因素方差分析过程本质为多总体均值检验, 分析步骤分为:

- 1 提出假设检验问题
- 2 构造检验统计量
- 3 做出统计决策等

单因素方差分析: 统计模型及假设检验

■ 假设检验问题 (一)

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

$$H_1 : \mu_i (i = 1, 2, \cdots, k) \text{ 不全相等}$$

■ 统计模型 (一)

$$x_{ij} = \mu_i + \epsilon_{ij},$$

ϵ_{ij} : 处理 i 的第 j 个观测值的随机误差, 且 $\epsilon_{ij} \sim N(0, \sigma^2)$

三个基本假定

- 1 每个总体都服从正态分布 $N(\mu_i, \sigma_i^2)$
- 2 各总体方差相等: $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$
- 3 观测值独立

单因素方差分析: 统计模型及假设检验

- μ : 不考虑处理影响时全部观测数据的总均值
- $\mu_i - \mu = \alpha_i$: 第 i 个处理效应均值与总均值之差, 表示处理为 i 时对观测值的附加效应
- 统计模型 (二)

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

处理 i 的第 j 个观测值 x_{ij} 等于不考虑处理影响时的总均值 μ 加上处理为 i 时的附加效应 α_i , 及其处理 i 的第 j 个观测值的随机误差 ϵ_{ij}

- 假设检验问题 (二)

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

$$H_1 : \alpha_i (i = 1, 2, \cdots, k) \text{ 不全为 } 0$$

单因素方差分析: 构造检验统计量

- 为检验 H_0 是否成立, 需要确定检验的统计量. 如何构造这一统计量呢?
- 误差分解: $SST = SSE + SSA$

单因素方差分析: 构造检验统计量

- μ_i 未知: $\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$
- μ 未知: $\bar{x} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{n}$
- $SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$
- $SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$
- $SSA = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$

单因素方差分析: 构造检验统计量

- 可以证明: $SST = SSE + SSA$
- 方差分析就是要分析数据的总误差中有没有处理误差
- 固定 SST , SSA 越大, SSE 越小, 即 SSA/SSE 越大, 越有理由认为存在处理出差, 即拒绝 H_0 认为因素对观测指标有影响
- 因此我们需要考虑 SSA/SSE 的统计分布

单因素方差分析: 构造检验统计量

在 H_0 成立时, 可以证明

$$1 \quad \frac{SST}{\sigma^2} = \sum_{i=1}^k \sum_{j=1}^{n_i} \left(\frac{x_{ij} - \bar{x}}{\sigma} \right)^2 \sim \chi^2(n-1)$$

$$2 \quad \frac{SSE}{\sigma^2} = \sum_{i=1}^k \sum_{j=1}^{n_i} \left(\frac{x_{ij} - \bar{x}_i}{\sigma} \right)^2 \sim \chi^2(n-k)$$

$$3 \quad \frac{SSA}{\sigma^2} = \sum_{i=1}^k n_i \left(\frac{\bar{x}_i - \bar{x}}{\sigma} \right)^2 \sim \chi^2(k-1)$$

■ $\frac{SSA}{\sigma^2}$ 和 $\frac{SSE}{\sigma^2}$ 互相独立

单因素方差分析: 构造检验统计量

H_0 为真时,

$$F = \frac{SSA/(k-1)}{SSE/(n-k)} \sim F(k-1, n-k)$$

- 离差平方和除以相应的自由度为均方误

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

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

- F 统计量: 组间均方误与组内均方误之比:

$$F = \frac{MSA}{MSE}$$

单因素方差分析: 作出统计决策

- 拒绝域 $W = \{F > c\}$

给定显著性水平 α ,

- $F > F_{\alpha}(k-1, n-k)$, 拒绝原假设, 表明所检验的因素对观测值有显著影响。
- $F < F_{\alpha}(k-1, n-k)$, 不能拒绝原假设。

单因素方差分析: 方差分析表

误差来源	平方和	自由度	均方	F
处理效应	$SSA = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$	$k - 1$	$MSA = SSA/(k-1)$	MSA/MSE
误差	$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$n - k$	$MSE = SSE/(n - k)$	
总	$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	$n - 1$		

单因素方差分析: 开篇案例

误差来源	平方和	自由度	均方	F	p 值
处理效应	$SSA = 1457$	3	$MSA = 485.5$	3.407	0.0388
误差	$SSE = 2708$	19	$MSE = 142.5$		
总	$SST = 4165$	22			

R 中做方差分析

```
counts <- c(57, 66, 49, 40, 34, 53, 44,  
            68, 39, 29, 45, 56, 51,  
            31, 49, 21, 34, 40,  
            44, 51, 65, 77, 58)  
industries <- c(rep('零售业', 7), rep('旅游业', 6),  
               rep('航空业', 5), rep('家电制作业', 5))  
fit <- aov(counts~industries)  
summary(fit)
```


单因素方差分析: 多重比较

- 开篇案例中结论: 不同行业被投诉次数的均值不完全相同. 但究竟哪些均值之间不相等? 这种差异到底出现在哪些行业之间?
- 多重比较方法 (Multiple comparison procedures): 通过对总体均值之间的配对比较进一步检验到底哪些均值之间存在差异
- 这里重点介绍 Fisher 提出的最小显著差异方法 (Least significant difference, LSD)

单因素方差分析: 多重比较

- 提出假设 $H_0: \mu_i = \mu_j$

- 计算检验统计量, H_0 成立时,

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim t(n - k)$$

- 写出拒绝域, 作出决策

- $LSD = t_{\alpha/2}(n - k) \sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$, 当 $|\bar{x}_i - \bar{x}_j| > LSD$ 时拒绝 H_0

- 思考开篇案例

Outline

- 1 一个例子
- 2 方差分析的本质及误差分解
- 3 单因素方差分析
- 4 双因素方差分析 (拓展内容)

双因素方差分析 (拓展内容)

双因素方差分析分析两个因素 (行因素和列因素) 对试验的影响 (自学内容)

【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？($\alpha=0.05$)

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

小结

- 了解什么是方差分析及本质 (多总体均值检验)
- 理解并熟悉掌握方差分析的基本思想及一般原理 (误差分解原理)
- 掌握单因素方差分析统计模型、假设检验、一般步骤及其 R 实现等
- 了解方差分析中关系强度的测量及其多重比较方法 (LSD)
- 自学：方差分析基本假定的检验、双因素方差分析等拓展内容

作业

习题 8.1: 1, 5