

第七讲 方差分析与实验设计

7.1 回顾假设检验问题

在参数检验问题中，常见的检验内容有：

例1： 某炼铁厂铁水含碳量： $X \sim N(4.55, 0.108^2)$

记 $\mu_0 = 4.55$ 。现改变工艺条件。检测5炉铁水，其含碳量为：

4.28, 4.40, 4.42, 4.35, 4.37

问工艺条件改变后，铁水含碳量是否改变？

$$X \sim N(\mu, \sigma^2) \leftarrow (X_1, \dots, X_n); \quad H_0 : \mu = \mu_0 ?$$

记: $\mu_0 = 4.55$

H_0 : $\mu = \mu_0$ (所观察到的现象是随机误差造成的)

H_1 : $\mu \neq \mu_0$ (所观察到的现象是真实的)

构造“检验统计量”: $Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2 / n}} \stackrel{H_0}{\sim} N(0,1)$

选择检验水平: $\alpha (= 0.05)$ $P\{|Z| \geq 1.96\} = 0.05$

$$|Z| = \left| \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right| = \left| \frac{4.364 - 4.55}{0.108 / \sqrt{5}} \right| = 3.85 > 1.96$$

决策: 拒绝 H_0 , 认为 $\mu \neq \mu_0$

例2:

某家庭日用食品商店在六个月内作了两次调查，以了解家庭每月平均消费量有无变化。结果如下

第一次: $n_1 = 200$, $\bar{x}_1 = 4.91$, $s_1^2 = 0.4$

第二次: $n_2 = 250$, $\bar{x}_2 = 4.97$, $s_2^2 = 0.5$

$$\left. \begin{array}{l} X \sim N(\mu_1, \sigma_1^2) \leftarrow (X_1, \dots, X_{n_1}) \\ Y \sim N(\mu_2, \sigma_2^2) \leftarrow (Y_1, \dots, Y_{n_2}) \end{array} \right\} H_0 : \mu_1 = \mu_2 ?$$

$$(1) H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$$(2) \text{ 统计量: } Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \stackrel{H_0}{\sim} N(0,1)$$

$$(3) \alpha = 0.05 \Rightarrow \pm z_{\alpha/2} = \pm 1.96$$

$$(4) z = \frac{4.91 - 4.97}{\sqrt{\frac{0.4}{200} + \frac{0.5}{250}}} = -0.9487 > -1.96$$

(5) 不能拒绝 H_0

7.2 方差分析的基本概念

方差分析的一般问题

检验多个总体均值是否相等的统计方法，称为方差分析（analysis of variance, 缩写**ANOVA**）

$$\left. \begin{array}{l} X_1 \sim N(\mu_1, \sigma_1^2) \leftarrow (X_{11}, \dots, X_{1n_1}) \\ X_2 \sim N(\mu_2, \sigma_2^2) \leftarrow (X_{21}, \dots, X_{2n_2}) \\ \vdots \\ X_r \sim N(\mu_r, \sigma_r^2) \leftarrow (X_{r1}, \dots, X_{r,n_r}) \end{array} \right\} H_0 : \mu_1 = \mu_2 = \dots = \mu_r ?$$

例3：某果酱制造企业希望了解哪种包装的罐头更受消费者欢迎，以确定其包装策略。传统的包装方法是用罐头。市场部经理则提议增添两种新包装：玻璃瓶、塑料瓶。为了避免大量生产的危险，公司接到该建议后，随机选定了三家所在区域与规模都近似的超市进行实验，分别销售一种包装的罐头（采用随机的方法决定哪家超市销售哪一种包装），实验期为4周。

要检验的对象（因素，factor）：包装方式

因素的不同表现（水平，处理 treatment）：

三种不同的包装形式

H_0 : 三种包装的平均销量相同

$$\mu_1 = \mu_2 = \mu_3$$

重复 次数 (i)	三种处理		
	罐头	玻璃	塑料
1	30	42	18
2	40	46	26
3	18	38	40
4	24	50	36
$x_{\bullet j} = \sum_{i=1}^4 x_{ij}$	112	176	120
\bar{x}_j	28	44	30

显然，三种包装的平均销售量不同。但这种差别是随机因素造成的，还是从总体上，它们确实存在明显的差异呢？

$$H_0: \mu_1 = \mu_2 = \mu_3$$

方差分析的基本工具： F-检验

复习： F-检验

$$X \sim N(\mu_1, \sigma_1^2) \leftarrow (X_1, X_2, \dots, X_{n_1}) \quad S_1^2$$

$$Y \sim N(\mu_2, \sigma_2^2) \leftarrow (Y_1, Y_2, \dots, Y_{n_2}) \quad S_2^2$$

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1), \text{ 若 } \sigma_1^2 = \sigma_2^2, X \text{ 与 } Y \text{ 相互独立}$$

7.3 单因素方差分析

例 3：果酱包装策略：罐头、玻璃瓶、塑料瓶。

随机选定了 3 家所在区域与规模都近似的超市进行实验，每家超市分别销售一种包装的罐头，实验期为4周。

实验单元 (experimental unit)

接受“处理”的对象或实体 (3个超市)

随机化设计：

将 k 种“处理”随机地指派给各个实验单元，
在每一个实验单元 (i), 重复 n_i 次实验

实验次数	罐头	玻璃	塑料
1	30	42	18
2	40	46	26
3	18	38	40
4	24	50	36

重复 次数 (i)	三种处理			每周销售
	罐头	玻璃	塑料	$x_{i\bullet} = \sum_{j=1}^3 x_{ij}$
1	30	42	18	90
2	40	46	26	112
3	18	38	40	96
4	24	50	36	110
$x_{\bullet j} = \sum_{i=1}^4 x_{ij}$	112	176	120	$x_{\bullet\bullet} = 408$
\bar{x}_j	28	44	30	$\bar{x} = 34$

三种包装的平均销售量不同。问题是这种差别是随机因素造成的，还是从总体上，它们确实存在明显的差异呢？

$$H_0: \mu_1 = \mu_2 = \mu_3$$

检验多个总体均值是否相同

二. 基本方法

1. 问题: X_i 正态分布, 同方差, 相互独立

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r ?$$

$$X_1 \sim N(\mu_1, \sigma^2) \leftarrow (X_{11}, \cdots, X_{1n_1}): \quad \bar{X}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} X_{1j}$$

$$X_2 \sim N(\mu_2, \sigma^2) \leftarrow (X_{21}, \cdots, X_{2n_2}): \quad \bar{X}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j}$$

\vdots

$$X_r \sim N(\mu_r, \sigma^2) \leftarrow (X_{r1}, \cdots, X_{rn_r}): \quad \bar{X}_r = \frac{1}{n_r} \sum_{j=1}^{n_r} X_{rj}$$

总的样本容量 n : $n = (n_1 + n_2 + \cdots + n_r) = \sum_{i=1}^r n_i$

总的样本均值: $\bar{X} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^r n_i \bar{X}_i$

各样本均值的
加权求和

2. 计算数据的总变差: $Q = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$

变差分解:

$$(X_{ij} - \bar{X}) = (X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})$$

两边平方, 求和:

$$\begin{aligned} \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 \\ &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2 \end{aligned}$$

• 注意:

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}) = \sum_{i=1}^r (\bar{X}_i - \bar{X}) \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i) = 0$$

变差分解的结果:

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2$$

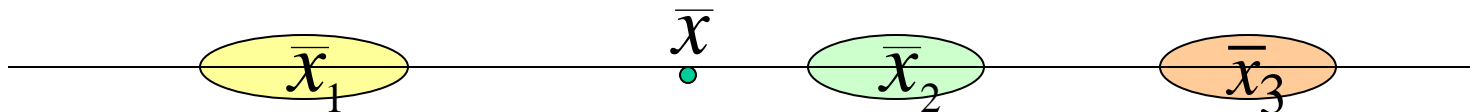
总变差 = 随机抽样误差 + 系统变差

(组内变差平方和) + (组间变差平方和)

$$Q = Q_1 + Q_2$$

$$\text{自由度 } (n-1) = (n-r) + (r-1)$$

$$\text{方差 } S^2 = Q / (n-1) \quad S_1^2 = Q_1 / (n-r) \quad S_2^2 = Q_2 / (r-1)$$



S_1^2 : 每组观测数据的方差（随机误差和）

S_2^2 : 每个总体的样本均值之间的差异

比较 两种方差的大小: S_2^2 是否明显大于 S_1^2

检验统计量:
$$F = \frac{S_2^2}{S_1^2} = \frac{Q_2 / (r-1)}{Q_1 / (n-r)} \stackrel{\sigma_1=\sigma_2}{\sim} F(r-1, n-r)$$

方差分析表

方差来源	平方和	自由度	方差	F值
因子影响	$Q_2 = \sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2$	$r-1$	$S_2^2 = Q_2 / (r-1)$	S_2^2 / S_1^2
随机误差	$Q_1 = \sum_{i=1}^r \sum_j^{n_i} (X_{ij} - \bar{X}_i)^2$	$n-r$	$S_1^2 = Q_1 / (n-r)$	
总和	$Q = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$	$n-1$	$S^2 = Q / (n-1)$	

例题的计算结果

方差分析表 ($r=3, n=12$)

方差来源	平方和	自由度	方差	F值
因子影响	608	$3-1=2$	304	4.275
误差	640	$12-3=9$	71.11	
总和	1248	$12-1=11$		

取 $\alpha=0.05$, 查表: $F_{0.05}(2, 9) = 4.26$

因为 $F = 4.275 > 4.26$

拒绝 H_0 , 即包装对销量的影响是比较明显的。

7.4 实验设计的基本原则

一、一个你熟悉的故事

例子：20世纪50年代，美国公共卫生总署组织小儿麻痹症疫苗实验：200000个受试者，和相同数目的对照组。在对照组中，有138个孩子患病；在实验组中，有56个孩子患病。用随机性是否能解释这个差别？

检验问题：两类孩子的患病率是否相等

工作目的：检验疫苗与患病率之间的**因果关系**

实验结果的比较（1954年）

随机对照双盲实验			NFIP研究		
	人数	比率 (10万分之)		人数	比率 (10万分之)
处理组	200000	28	2年级(接种)	22500	25
对照组	200000	71	1、3年级（对照）	725000	54
不同意	350000	46	2年级不同意	125000	44

（1）从随机对照双盲实验可以看出，处理组的患病率明显低于对照组；（71-28）

（2）NFIP对照组包含一部分父母不同意的儿童，所以患病率较低。结果使得两组患病率之差下降（54-25）。

（3）在NFIP实验中，其结果还可能受到儿童在2年级等因素的影响。而调研者缺乏足够的资料以估计其对实验结果的影响。

实验设计

工作目的：检验变量之间的因果关系

科学研究方法：先对某一过程或系统有一些猜想，
然后通过实验产生有关的数据，并对猜想进行验证。

实验设计(Experimental Design)：

为收集样本数据所进行的计划

实验设计规则：控制某一情形的所有相关方面，
操纵少数感兴趣的变量，然后观察实验结果。

7.4 实验设计的基本原则

(一) 小儿麻痹症疫苗实验

实验方案1：如果在1954年对大量儿童进行接种实验。结果发现1954年的发病率确实比1953年急剧下降。**能否证实疫苗确实有显著效果？**

小儿麻痹症是一种每年发病率变化很大的流行病。

事实上，1952年大约有60000个病历，而1953年仅有其一半。

所以存在两种可能：

- 疫苗是有效的
- 当年没有流行此病

实验方案2：在同一年份中，一部分儿童接种疫苗，而另一部分儿童不接种疫苗。当然，只有取得父母同意的儿童才能接种疫苗。一个容易操作的方案是：那些取得父母同意的儿童接种疫苗，其余儿童组成对照组。

问题：两组孩子的家庭背景不同。高收入家庭的父母常常比低收入家庭的父母更赞同接种疫苗；而高收入家庭的儿童更容易遭受小儿麻痹疫病的伤害。

生活在卫生条件比较差的儿童，在童年早期尚受到来自母亲抗体保护时，就有可能轻微感染过这种病。

必须避免混淆：两组之间的任何差异只能归因于处理不同，而不是其他原因。

实验方案3：小儿麻痹症全国基金会（NFIP）的方案

所有小学2年级的并取得父母同意的儿童接种疫苗，而1年级和3年级儿童组成对照组。

该实验设计的缺陷：

实验方案4：随机对照双盲实验

(1) 实验组和对照组来自同一总体（家庭收入、儿童一般健康状况、性格以及社会习惯基本相似）：

(2) 以50%对50%的机会，随机分配每一个儿童到处理组或者是对照组。（保证这两个组在一些重要变量上的取值分配是相近的）

(3) 使用安慰剂：给对照组儿童注射盐溶液（实验对象不知道是在处理组还是实验组，避免精神力量作用）

(4) 诊断医生（评估反映的人）不知道孩子是属于哪一组的。由于小儿麻痹病的诊断比较复杂，避免医生=在诊断时可能受先验信息的影响。（双盲实验）

（二）安妥明实验

评估某种药品对冠心病的治疗效果。在对心脏有问题的男子（实验对象）中，随机指派1103名病人到药物组，2789名到对照组。跟踪观察5年。

	安妥明		安慰剂	
	人数	死亡率	人数	死亡率
坚持者	708	15%	1813	15%
不坚持者	357	25%	882	28%
整个组	1103	20%	2789	21%

结论：（1）安妥明没有显著效果

（2）坚持者与不坚持者除所服药物之外，在其他方面有所不同。（他们更关心自己的健康，在总的方面对自己更加照顾。）

实验设计的特点：控制某一事物的所有相关方面，操纵少数感兴趣的变量，然后观察实验结果。

（三）完全随机化设计——单因素方差分析

某产品开发工程师考虑能使一种新的合成纤维的抗拉强度增大的方案（这种合成纤维织出的布是用来缝制男士衬衫的）。他推测适度提高棉花在纤维中含量，有可能增加抗拉强度。而且根据以往经验知道，棉花含量应该在10%到40%之间。他决定检验棉花百分率为5个水平的样品，这5个水平分别是：15%，20%，25%，30%，35%。他还决定，使用一台抗拉强度测试机，对每个水平试验5个样品（ $n=5 \times 5=25$ ）。

棉花百分率 %	试验号数				
15	1	2	3	4	5
20	6	7	8	9	10
25	11	12	13	14	15
30	16	17	18	19	20
35	21	22	23	24	25

问题：如何确定这 25个 试验的次序（是否可以先做5个含量15%的样品，然后再做5个含量20%的样品，……？）

防止干扰变量对实验结果的污染：

抗拉强度检测机显示出某种热效应，运行的时间越长，读出的抗拉强度值越低。按照随机化的次序进行试验：

试验序列	试验号数	棉花百分率	试验序列	试验号数	棉花百分率
1	8	20	14	7	20
2	18	30	15	1	15
3	10	20	16	24	35
4	23	35	17	21	35
5	17	30	18	11	25
6	5	15	19	2	15
7	14	25	20	13	25
8	6	20	21	22	35
9	15	25	22	16	30
10	20	30	23	25	35
11	9	20	24	19	30
12	4	15	25	3	15
13	12	25			

完全随机化设计的特点： 试验以随机顺序的方式进行，使得每个处理所处的环境都尽可能均匀。

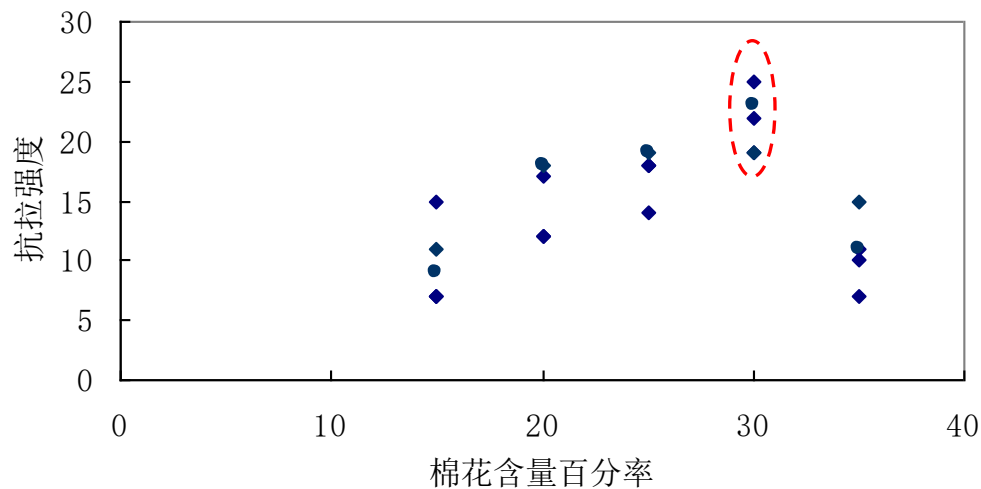
抗拉强度的实验数据

棉花百分率	观察值					总和	平均值
	1	2	3	4	5		
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						376	15.04

单因素方差分析

SUMMARY

组	观测数	求和	平均	方差
15	5	49	9.8	11.2
20	5	77	15.4	9.8
25	5	88	17.6	4.3
30	5	108	21.6	6.8
35	5	54	10.8	8.2



F	P-value	F crit
14.75682	9.13E-06	2.866081

结论： 差异显著。 棉花含量30%时，抗拉强度最高

阅读与练习

单因素方差分析Excel软件应用

无交互作用双因素方差分析Excel软件应用

例1：某果酱制造企业希望了解哪种包装的罐头更受消费者欢迎，以确定其包装策略。传统的包装方法是用罐头。市场部经理则提议增添两种新包装：玻璃瓶、塑料瓶。为了避免大量生产的危险，公司接到该建议后，随机选定了三家所在区域与规模都近似的超市进行实验，分别销售一种包装的罐头（采用随机的方法决定哪家超市销售哪一种包装），实验期为4周。

要检验的对象（因素，factor）： 包装方式

因素的不同表现（水平，处理treatment）：

三种不同的包装形式

H_0 ：三种包装的平均销量相同

重复 次数 (i)	三种处理			每周销售
	罐头	玻璃	塑料	$x_{i\bullet} = \sum_{j=1}^3 x_{ij}$
1	30	42	18	90
2	40	46	26	112
3	18	38	40	96
4	24	50	36	110
$x_{\bullet j} = \sum_{i=1}^4 x_{ij}$	112	176	120	$x_{..} = 408$
\bar{x}_j	28	44	30	$\bar{x} = 34$

三种包装的平均销售量不同。问题是这种差别是随机因素造成的，还是从总体上，它们确实存在明显的差异呢？

$$H_0: \mu_1 = \mu_2 = \mu_3$$

检验多个总体均值是否相同

Excel: 工具 → 数据分析 → 单因素方差分析

数据: 课堂案例数据——方差分析——包装方式

分组方式: 列
标志位于第一行

Microsoft Excel - 包装方式				
文件 (F) 编辑 (E) 视图 (V) 插入 (I)				
M16 fx				
	A	B	C	
1	罐头	玻璃	塑料	
2	30	42	18	
3	40	46	26	
4	18	38	40	
5	24	50	36	
6				

Microsoft Excel - 包装方式

文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数据(D) 窗口(W)



M16

f_x

	A	B	C	D	E	F	G
1	罐头	玻璃	塑料				
2	30	42	18				
3	40						
4	18						
5	24						
6							
7							
8							
9							
10							
11							
12							
13							

数据分析

分析工具(A)

方差分析：单因素方差分析
方差分析：可重复双因素分析
方差分析：无重复双因素分析
相关系数
协方差
描述统计
指数平滑
F-检验 双样本方差
傅利叶分析
直方图

确定

取消

帮助(H)

Microsoft Excel - 包装方式

文件(F) 编辑(E) 视图(V) 插入(I) 格式(O) 工具(T) 数据(D) 窗口(W) 帮助

常用 格式 数据 窗口 帮助

M16							
-----	--	--	--	--	--	--	--

	A	B	C	D	E	F	G
1	罐头	玻璃	塑料				
2	30						
3	40						
4	18						
5	24						

方差分析：单因素方差分析

输入

输入区域(I):

分组方式:
☒ 列(C)
☐ 行(R)

☒ 标志位于第一行(L)

α (A):

输出选项

☐ 输出区域(O):

☒ 新工作表组(P):

☐ 新工作簿(W)

确定

取消

帮助(H)

6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							

方差分析：单因素方差分析						
SUMMARY						
组	计数	求和	平均	方差		
罐头	4	112	28	88		
玻璃	4	176	44	26.66667		
塑料	4	120	30	98.66667		
方差分析						
差异源	SS	df	MS	F	P-value	F crit
组间	608	2	304	4.275	0.049527	4.256492
组内	640	9	71.11111			
总计	1248	11				

关系强度的测量

测量自变量与因变量之间的关系强度：

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

方差分析：单因素方差分析						
SUMMARY						
组	计数	求和	平均	方差		
罐头	4	112	28	88		
玻璃	4	176	44	26.66667		
塑料	4	120	30	98.66667		
方差分析						
差异源	SS	df	MS	F	P-value	F crit
组间	608	2	304	4.275	0.049527	4.256492
组内	640	9	71.11111			
总计	1248	11				

包装方式对销量的影响强度： $R^2 = \frac{608}{1248} = 0.487$

案例

早期案例：17世纪初,海上长期航行的水手常患坏血病。英国海军试图发现坏血病的起因：一艘船上的水手每天喝柑橘汁,另外三艘船的水手则没有柑橘汁供应。

检验问题：两类船水手的患病率是否相等

实验缺陷： 如果随机决定哪一个水手喝或不喝柑橘汁，而不是按船来分，可以消除有关船的其他因素的影响，结果会更有说服力。）

例2：测量赠品的促销效果

某百货公司欲在**12月**采用赠品政策来促销某品牌的毛衣。

采用赠品政策前的销售量： 100件

采用赠品政策后的销售量： 130件

问题： 赠品效果如何 $130-100=30$

没有考虑12月份圣诞节和新年节日的影响。这期间销售量较往日高，不一定是赠品政策的影响。

	实验组	对照组
事前	销量(有赠品) 100	销量(无赠品) 100
事后	销量(有赠品) 130	销量(无赠品) 110
变动	30	10