《传播统计学》

单因素方差分析

教师：林志良
邮箱：linzhl@nfu.edu.cn
个人网站：www.zhilianglin.com

目录

单因素方差分析介绍

- **作用**：比较多组样本的均值
  - **适用条件**：自变量-多类别变量；因变量-数值型变量
  - 例如：
    - 文传各专业（汉语言/新闻/网新）同学社交媒体使用时间差异
    - 不同生源地（珠三角/非珠三角省内/省外）同学英语口语能力差异
- 方差分析有多种类型，这里我们只学习单因素方差分析（One-way ANOVA）

Analysis of Variance

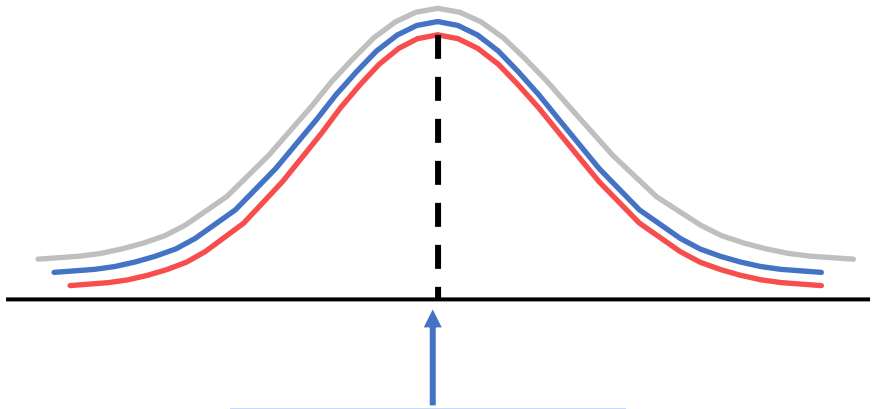
单因素方差分析介绍

假设

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

H_1: 不是所有的μ_i 都相等的

- **如果H_0成立：**



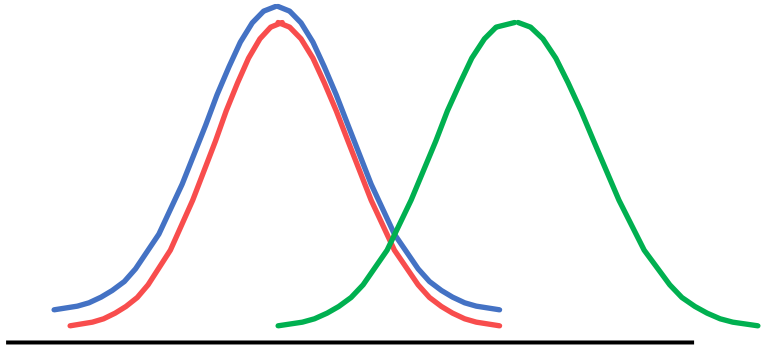
单因素方差分析介绍

假设

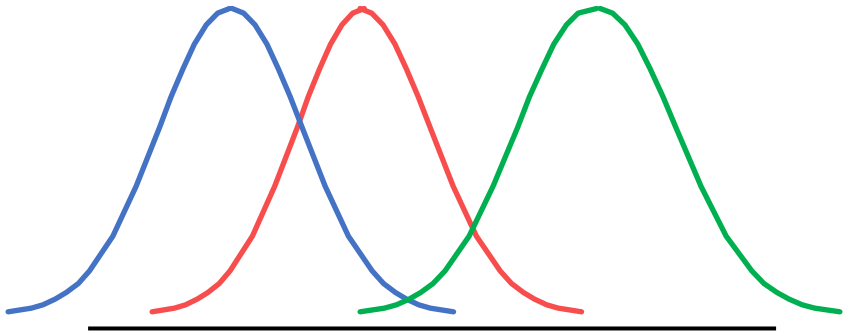
$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$$

H_1: 不是所有的μ_i 都相等的

- **如果H_0不成立：**



或



$\mu_1 = \mu_2 \neq \mu_3$

$\mu_1 \neq \mu_2 \neq \mu_3$

单因素方差分析介绍

F检验的思想

首先将总变异（$SS_{总}$）分解为组间变异（$SS_{组间}$）和组内变异（$SS_{组内}$），然后比较两者的平均变异 $MS_{组间}$ 和$MS_{组内}$，比较时采用两者的比值F值，即：

$$F = \frac{MS_{组间}}{MS_{组内}}$$

解构变异

总变异=组间变异+组内变异

$$SS_{总} = SS_{组间} + SS_{组内}$$

$SS_{总}$	SST = Total Sum of Squares (Total variation)	总平方和 (总变异)
$SS_{组间}$	SSA = Sum of Squares Among Groups (Among-group variation)	组间平方和 (组间变异)
$SS_{组内}$	SSW = Sum of Squares Within Groups (Within-group variation)	组内平方和 (组内变异)

解构变异

总变异

$$SS_{总} = SS_{组间} + SS_{组内}$$

$$SS_{总} = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(X_{ij} - \overline{\overline{X}})^2$$

其中，

$SS_{总}$ = 总平方和

c = 组别数

n_i = 第i组的数据量

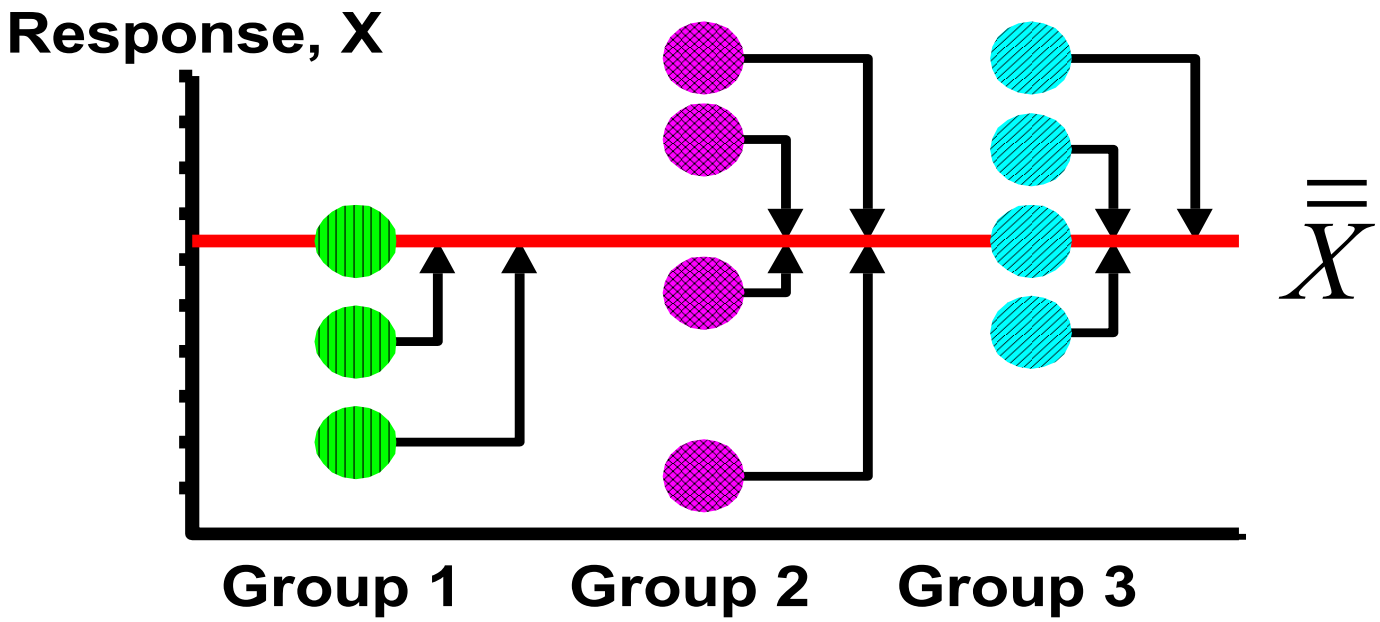
X_{ij} = 第i组的第j个观测值

$\overline{\overline{X}}$ = 所有数值的平均值

解构变异

总变异

$$SS_{总} = (X_{11} - \overline{\overline{X}})^2 + (X_{12} - \overline{\overline{X}})^2 + \cdots + (X_{cn_c} - \overline{\overline{X}})^2$$



解构变异

组间变异

$$SS_{总} = SS_{组间} + SS_{组内}$$

$$SS_{组间} = \sum_{i=1}^{c} n_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2$$

其中，

$SS_{组间}$ = 组间平方和

c = 组别数

n_i = 第i组的数据量

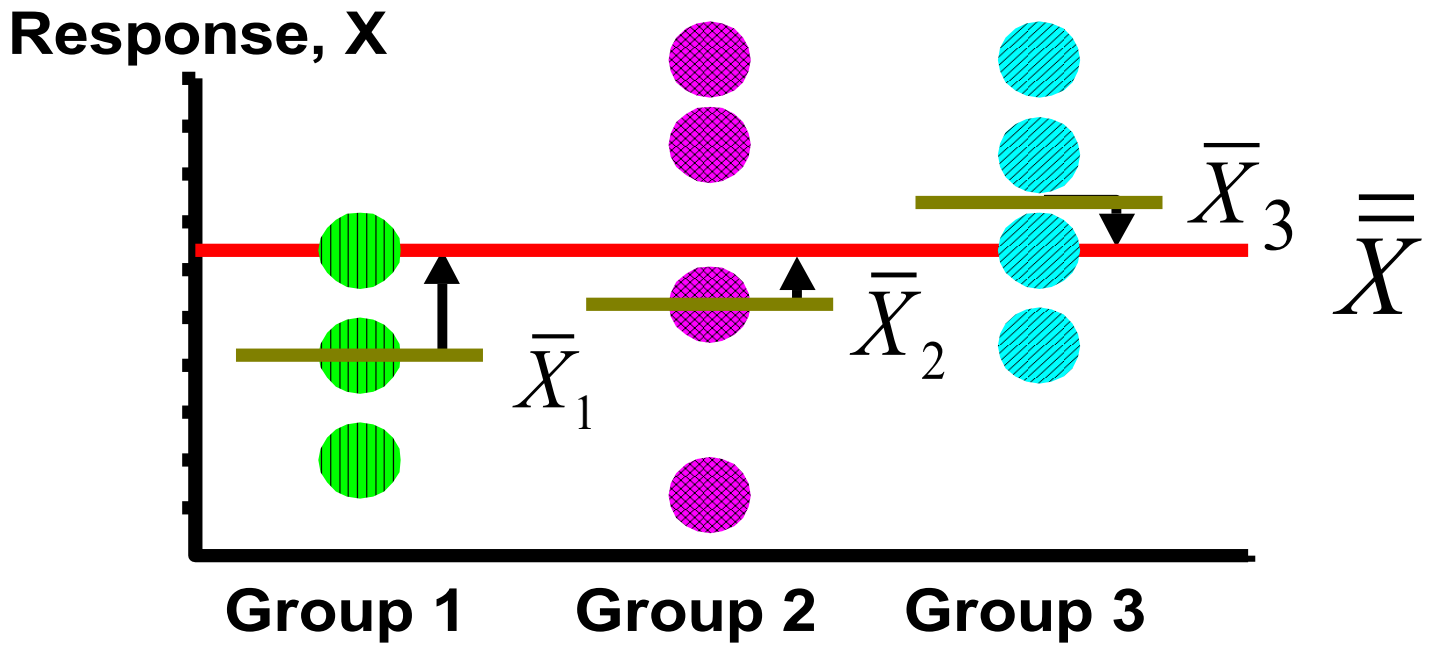
$\overline{X}_i$ =第i组的平均值

$\overline{\overline{X}}$ = 所有数值的平均值

解构变异

组间变异

$$SS_{组间} = n_1(\overline{X}_1 - \overline{\overline{X}})^2 + n_2(\overline{X}_2 - \overline{\overline{X}})^2 + \cdots + n_c(\overline{X}_c - \overline{\overline{X}})^2$$

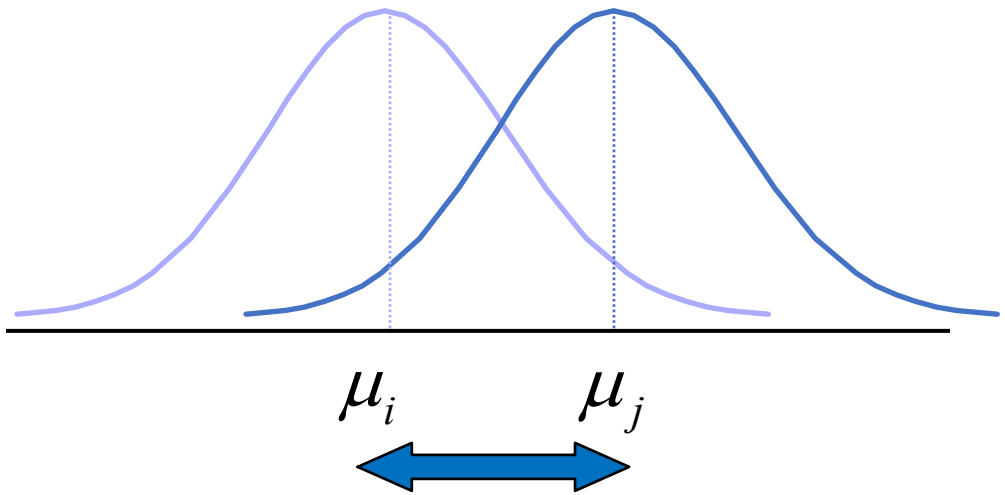


解构变异

组间平均变异

$$SS_{组间} = \sum_{i=1}^{c} n_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2$$

组间变异



$$MS_{组间} = \frac{SS_{组间}}{c - 1}$$

组间均方= 组间平方和/自由度

解构变异

组内变异

$$SS_{总} = SS_{组间} + SS_{组内}$$

$$SS_{组内} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$$

其中，

$SS_{组内}$ = 组内平方和

c = 组别数

n_i = 第i组的数据量

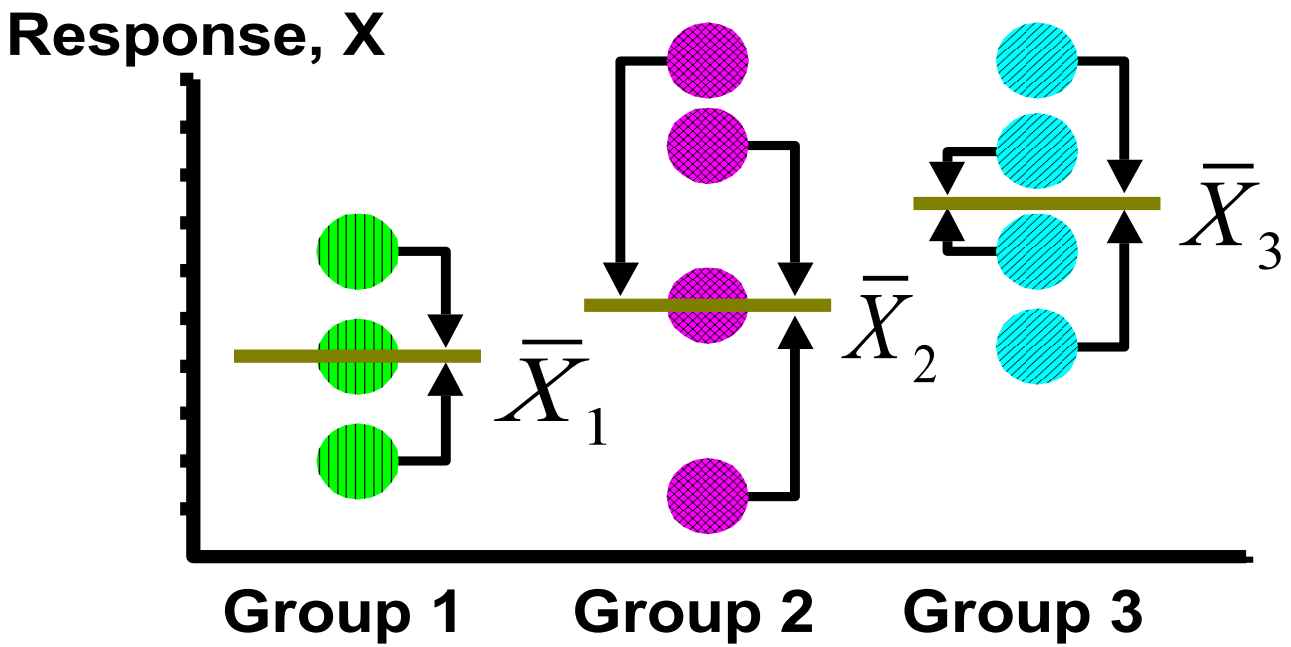
X_{ij} = 第i组的第j个观测值

$\overline{X}_i$ =第i组的平均值

解构变异

组内变异

$$SS_{组内} = (X_{11} - \overline{X}_1)^2 + (X_{12} - \overline{X}_1)^2 + \cdots + (X_{cn_c} - \overline{X}_c)^2$$

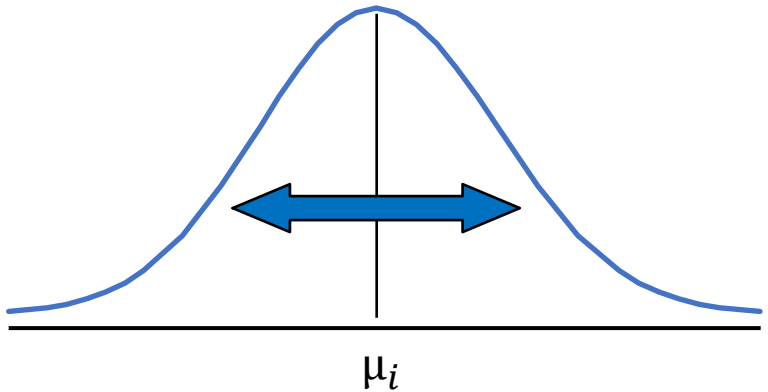


解构变异

组内平均变异

$$SS_{组内} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$$

计算每个组的内部变异，然后将各个组的组内变异数值加总



$$MS_{组内} = \frac{SS_{组内}}{n - c}$$

组内均方= 组内平方和/自由度

解构变异

均方

$$MS_{组间} = \frac{SS_{组间}}{df_{组间}}$$

Mean Square Among

($df_{组间} = c - 1$)

组间均方

$$MS_{组内} = \frac{SS_{组内}}{df_{组内}}$$

Mean Square Within

($df_{组内} = n - c$)

组内均方

$$MS_{总} = \frac{SS_{总}}{n - 1}$$

Mean Square Total

(d.f. = n-1)

总均方

解构变异

单因素方差分析表

变异 (Variance)	平方和 (Sum of Squares, SS)	自由度 (degree of freedom, df)	均方 (Mean Squares, MS)	F
组间	$SS_{组间}$	c -1	$MS_{组间}$	$\frac{MS_{组间}}{MS_{组内}}$
组内	$SS_{组内}$	n – c	$MS_{组内}$	
合计	$SS_{总}$	n - 1		

- c：组别数
- n：总样本量

- $df_{组间} = c - 1$
- $df_{组内} = n - c$

- $MS_{组间} = \frac{SS_{组间}}{df_{组间}}$
- $MS_{组内} = \frac{SS_{组内}}{df_{组内}}$

假设检验

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

H_1: 至少有两组总体均值不相等

- **检验统计量**

$$F = \frac{MS_{组间}}{MS_{组内}}$$

- **自由度**

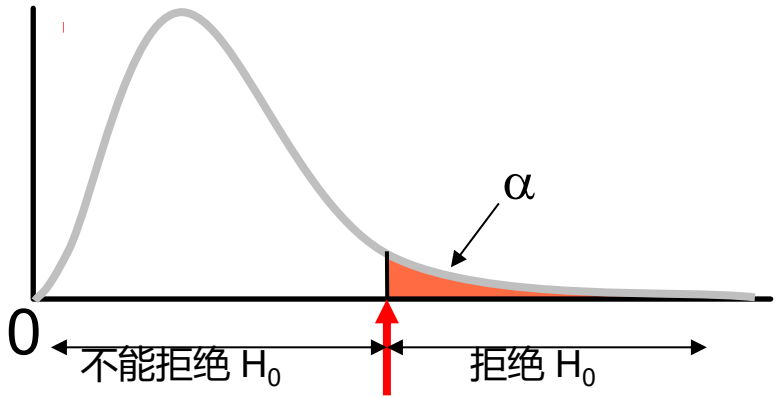
df_1 $df_{组间} = c - 1$ (c = 组别数)

df_2 $df_{组内} = n - c$ (n = 总样本量)

假设检验

假设检验决策:

如果F统计量 > F_α，则拒绝H_0，否则不能拒绝H_0



临界值F_α

注：在方差分析中我们只关心组间均方是否显著大于组内均方；如果组间均方小于组内均方，就无须检验其是否小到显著水平，因而我们进行的是单侧检验。

例题

为了了解三家高尔夫球俱乐部的会员高尔夫球的水平是否有差异，随机各从三家俱乐部挑出5位会员，测试他们打高尔夫球的平均距离。

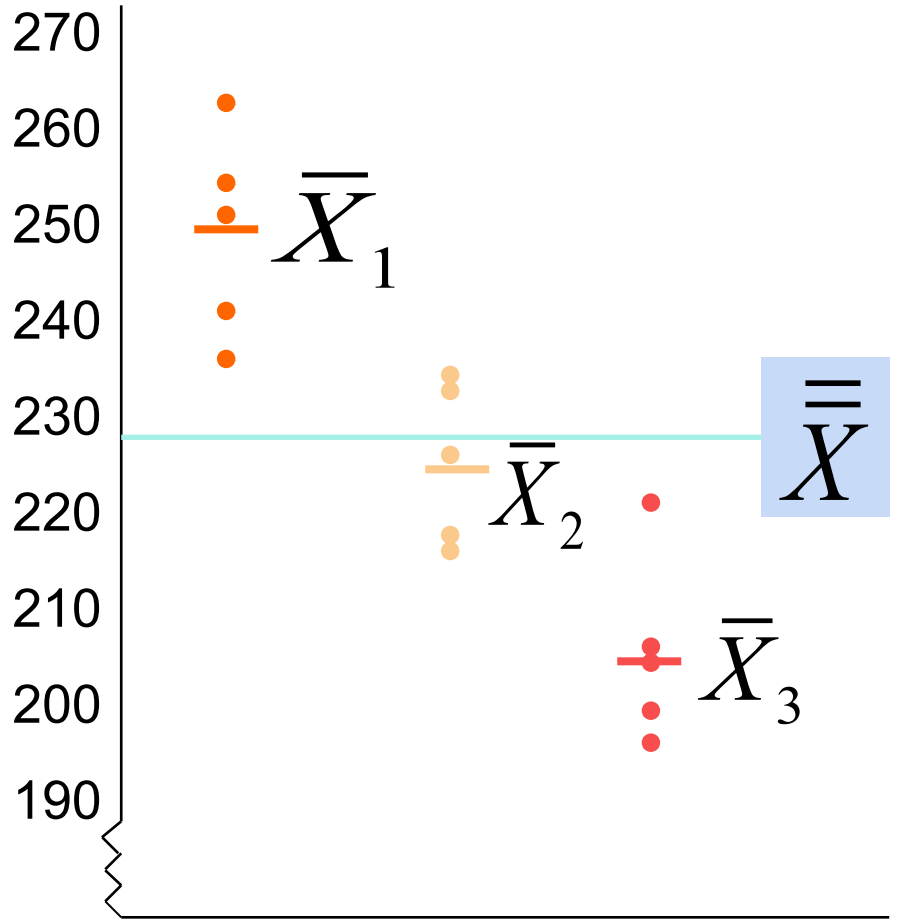
Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

例题

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$	$\bar{x}_2 = 226.0$	$\bar{x}_3 = 205.8$

$\bar{\bar{x}} = 227.0$



例题

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{X}_1 = 249.2$ $n_1 = 5$

$\bar{X}_2 = 226.0$ $n_2 = 5$

$\bar{X}_3 = 205.8$ $n_3 = 5$

$\bar{\bar{X}} = 227.0$ $n = 15$ $c = 3$

$SS_{组间} = 5\,(249.2 - 227)^2 + 5\,(226 - 227)^2 + 5\,(205.8 - 227)^2 = 4716.4$

$SS_{组内} = (254 - 249.2)^2 + (263 - 249.2)^2 + \cdots + (204 - 205.8)^2 = 1119.6$

$MS_{组间} = 4716.4 / (3-1) = 2358.2$

$MS_{组内} = 1119.6 / (15-3) = 93.3$

$$F = \frac{2358.2}{93.3} = 25.275$$

例题

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

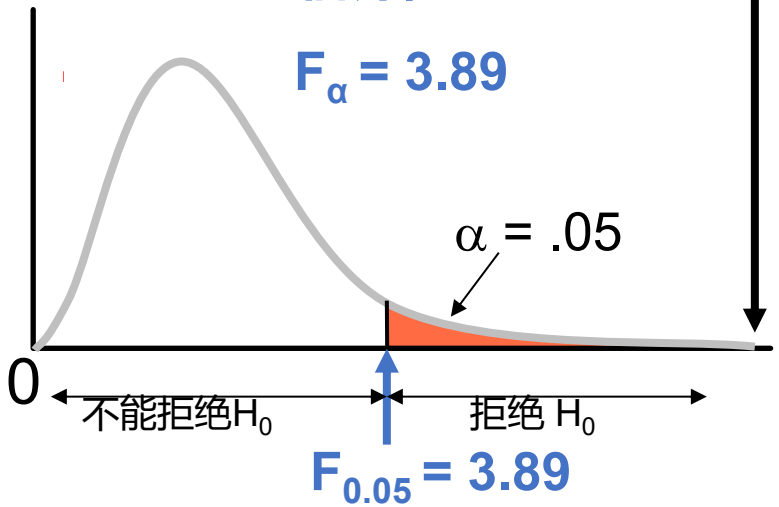
H_1:不是所有的μ_i 都相等

$\alpha = 0.05$

$df_1 = 2$　　$df_2 = 12$

临界值:

$F_\alpha = 3.89$



检验统计量:

$$F = \frac{MS_{组间}}{MS_{组内}} = \frac{2358.2}{93.3} = 25.275$$

假设检验决策:

在$\alpha = 0.05$的水平下拒绝H_0

结论:

三家俱乐部成员的球技并非完全一致

(至少有一家俱乐部的成员与其它俱乐部成员球技有显著差异)

例题

ANOVA

距离

	平方和	自由度	均方	F	显著性
组间	4716.400	2	2358.200	25.275	.000
组内	1119.600	12	93.300		
总计	5836.000	14			

练习

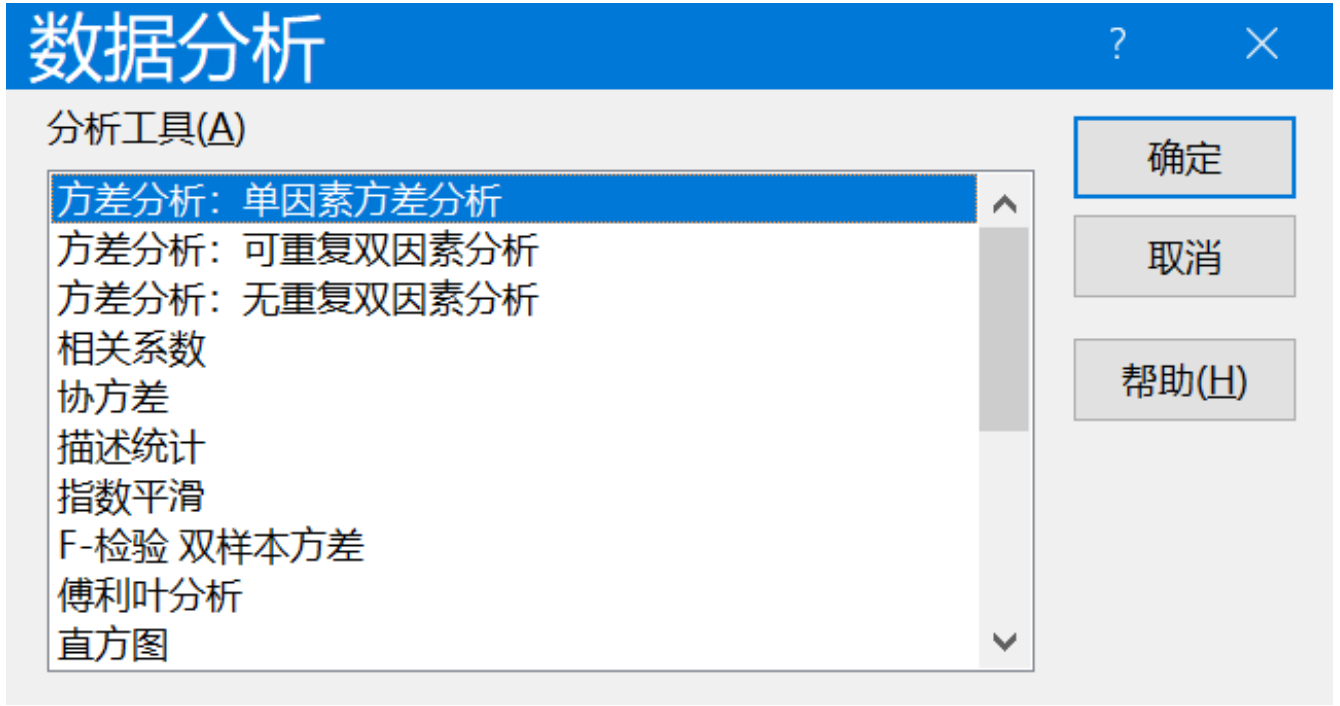
三个小组成员的成绩有差异吗?

第一组	第二组	第三组
82	79	83
81	80	84
82	80	83
82	81	85
83	80	85
$\bar{X}_1 = 82$	$\bar{X}_2 = 80$	$\bar{X}_3 = 84$

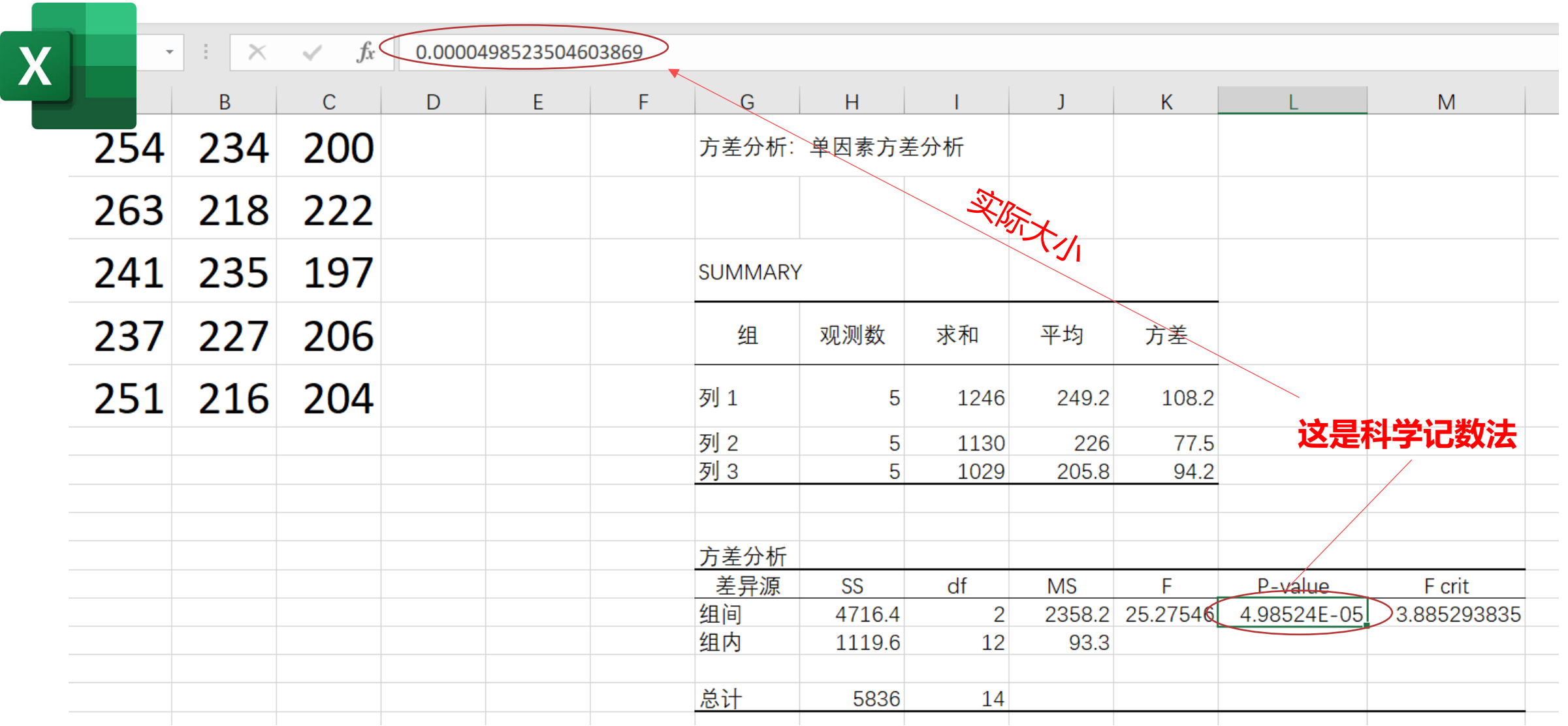
$\overline{\overline{X}} = 82$

软件实操

【工具】>【数据分析】>【方差分析：单因素方差分析】>选择数据范围



软件实操


0.0000498523504603869
254 234 200
263 218 222
241 235 197
237 227 206
251 216 204
方差分析：单因素方差分析
实际大小
SUMMARY
组 观测数 求和 平均 方差
列 1 5 1246 249.2 108.2
列 2 5 1130 226 77.5
列 3 5 1029 205.8 94.2
方差分析
差异源 SS df MS F P-value F crit
组间 4716.4 2 2358.2 25.27546 4.98524E-05 3.885293835
组内 1119.6 12 93.3
总计 5836 14
这是科学记数法

软件实操

将数据整理成右图的形式

距离	俱乐部
254.00	1.00
263.00	1.00
241.00	1.00
237.00	1.00
251.00	1.00
234.00	2.00
218.00	2.00
235.00	2.00
227.00	2.00
216.00	2.00
200.00	3.00
222.00	3.00
197.00	3.00
206.00	3.00
204.00	3.00

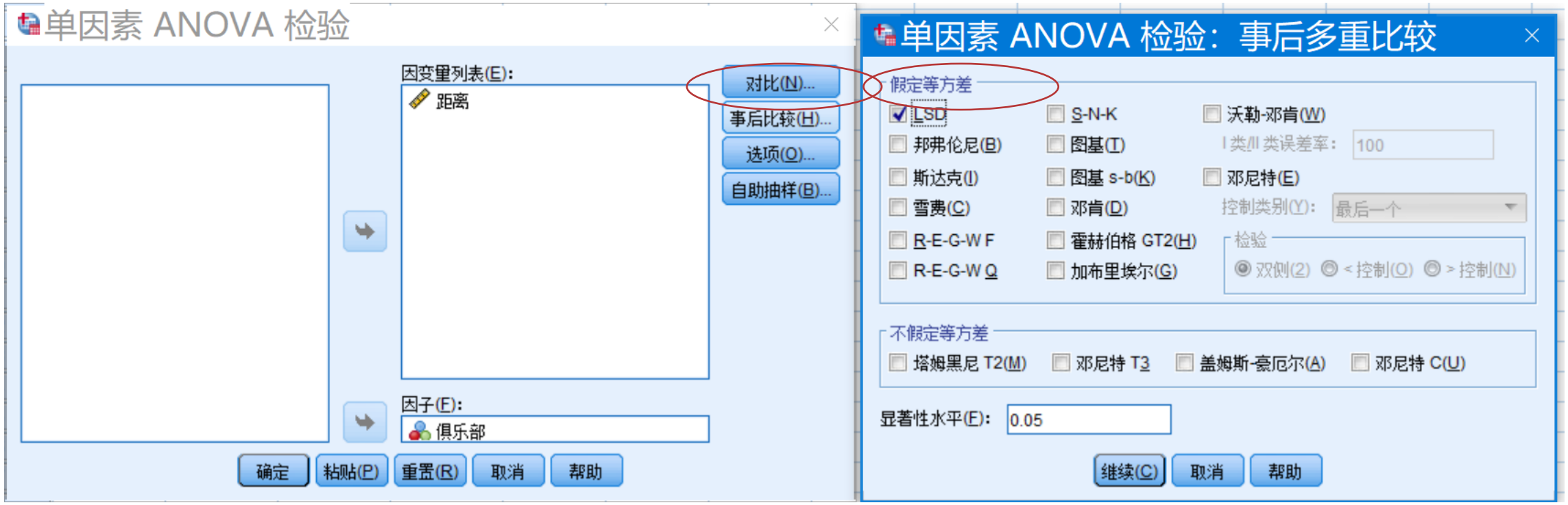
软件实操


分析(A)
图形(G)
实用程序(U)
扩展(X)
窗口(W)
帮助
报告(P)
描述统计(E)
贝叶斯统计信息(B)
表(B)
比较平均值(M)
一般线性模型(G)
广义线性模型(Z)
混合模型(X)
相关(C)
回归(R)
对数线性(O)
神经网络(W)
分类(F)
降维(D)
刻度(A)
非参数检验(N)
时间序列预测(T)
生存分析(S)
多重响应(U)
缺失值分析(Y)...
多重插补(T)
复杂抽样(L)
模拟(I)...
质量控制(Q)
空间和时间建模(S)...
直销(K)
平均值(M)...
单样本 T 检验(S)...
独立样本 T 检验...
成对样本 T 检验(P)...
单因素 ANOVA 检验...
Q2_选项3
Q2_选项4


单因素 ANOVA 检验
因变量列表(E):
距离
因子(F):
俱乐部
对比(N)...
事后比较(H)...
选项(O)...
自助抽样(B)...
确定
粘贴(P)
重置(R)
取消
帮助

软件实操

单因素分析中，只要有任意两组平均值不相等我们就可以拒绝原假设了，那么如果想知道具体是哪两组均值不相等呢？使用事后检验中的LSD方法（可以简单理解为做多个t检验）



软件实操


单因素 ANOVA 检验
因变量列表(E):
距离
因子(F):
俱乐部
对比(N)...
事后比较(H)...
选项(O)...
自助抽样(B)...
确定
粘贴(P)
重置(R)
取消
帮助
单因素 AN...
统计
描述(D)
固定和随机效应(F)
方差齐性检验(H)
布朗-福塞斯(B)
韦尔奇(W)
平均值图(M)
缺失值
按具体分析排除个案(A)
成列排除个案(L)
继续(C)
取消
帮助

软件实操

ANOVA

距离

	平方和	自由度	均方	F	显著性
组间	4716.400	2	2358.200	25.275	.000
组内	1119.600	12	93.300		
总计	5836.000	14			

软件实操

事后检验

多重比较

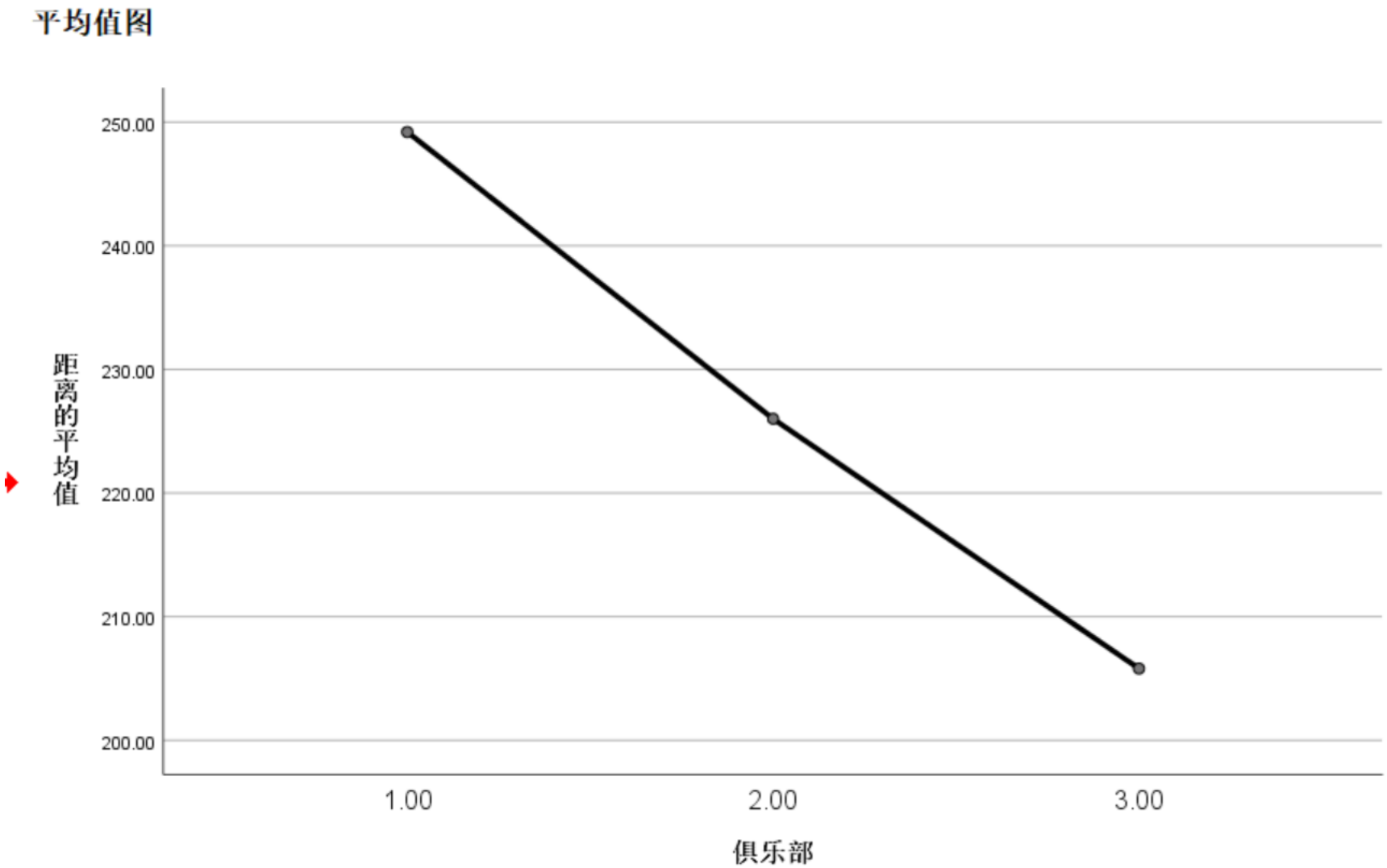
因变量: 距离

LSD

(I) 俱乐部	(J) 俱乐部	平均值差值 (I-J)	标准 错误	显著性	95% 置信区间 下限	95% 置信区间 上限
1.00	2.00	23.20000*	6.10901	.003	9.8896	36.5104
	3.00	43.40000*	6.10901	.000	30.0896	56.7104
2.00	1.00	-23.20000*	6.10901	.003	-36.5104	-9.8896
	3.00	20.20000*	6.10901	.006	6.8896	33.5104
3.00	1.00	-43.40000*	6.10901	.000	-56.7104	-30.0896
	2.00	-20.20000*	6.10901	.006	-33.5104	-6.8896

*. 平均值差值的显著性水平为 0.05。

软件实操


平均值图
250.00
240.00
230.00
220.00
210.00
200.00
距离的平均值
1.00
2.00
3.00
俱乐部


谢谢！