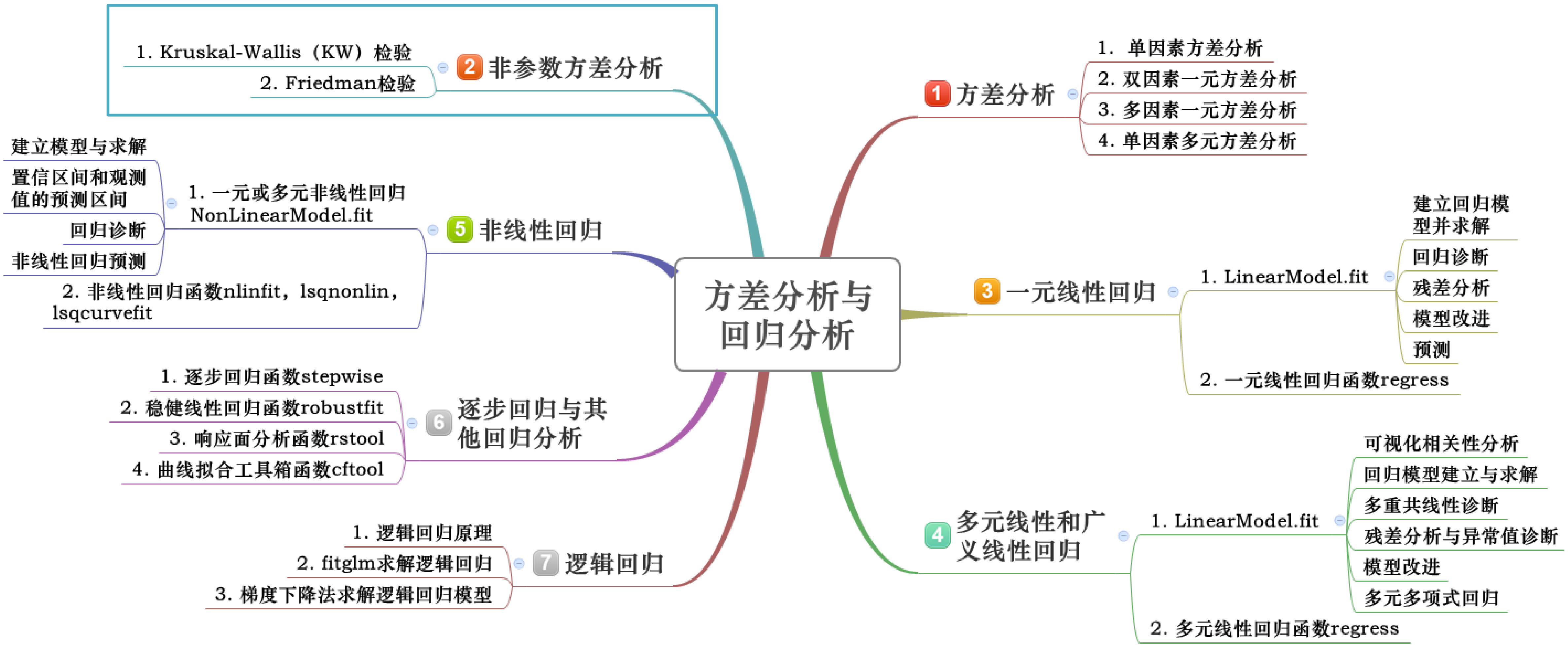
信陽師範學院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第11章 方差分析与回归分析

讲授人：牛言涛

日期：2020年4月14日

第11章 方差分析与回归分析知识点思维导图



11.2 非参数方差分析

- 前面介绍的方差分析均要求样本总体来自正态分布，并且这些正态总体应具有相同的方差，在这样的基本假定（正态性假定和方差齐次性假定）下检验总体均值是否相等，这属于参数检验。
- 当数据不满足正态性和方差齐次性假定时，参数检验可能会出现错误，此时应该采用基于秩的非参数检验，包括
  - 适用于完全随机化设计的单向秩次方差分析的Kruskal-Wallis（KW）检验（又称H检验）；首先，将多组样本数混合并按升序排序，求出各变量值的秩；然后，考察各组秩的均值是否存在显著差异。
  - 适用于随机化区组设计的双向秩次方差分析的Friedman检验。首先以行为单位将数据按升序排序，并求得各变量值在各自行中的秩；然后，分别计算各组样本下的秩总和与平均秩。多配对样本的Friedman检验适于对定距型数据的分析。

1. 非参数Kruskal-Wallis检验

- 多独立样本Kruskal-Wallis 检验的基本思想是：
  - 首先，将多组样本数混合并按升序排序，求出各变量值的秩；然后，考察各组秩的均值是否存在显著差异。
  - 如果各组秩的均值不存在显著差异，则认为多组数据充分混合，数值相差不大，多个总体的分布无显著差异；反之，如果各组秩的均值存在显著差异，则多组数据无法混合，有些组的数值普遍偏大，有些组的数值普遍偏小，可认为多个总体的分布存在显著差异，至少有一个样本不同于其他样本。
- 方差分析认为，各样本组秩的总变差一方面源于各样本组之间的差异（组间差），另一方面源于各样本组内的抽样误差（组内差）。如果各样本组秩的总变差的大部分可由组间差解释，则表明各样本组的总体分布存在显著差异；反之，如果各样本组秩的总变差的大部分不能由组间差解释，则表明各样本组的总体分布没有显著差异。

1. 非参数Kruskal-Wallis检验

例如：希望对A、B、C、D 四个城市的周岁儿童身高进行比较分析，采用独立抽样的方式获得四组独立样本，见表：

儿童	A	秩	B	秩	C	秩	D	秩
1	72	7	71	4	75	12.5	69	1
2	75	12.5	72	7	76	14.5	70	2
3	76	14.5	73	9	77	16	71	4
4	78	18	74	10.5	78	18	71	4
5	79	20	74	10.5	78	18	72	7
秩和	组A秩和	72	组B秩和	41	组C秩和	79	组D秩和	18
秩均	组A秩均	14.4	组B秩均	8.2	组C秩均	15.8	组D秩均	3.6

计算每个样本的秩总和。n1=5，n2=5，n3=5，n4=5，n=20，可得R1=72，R2=41，R3=79，R4=18，k=4，H0：四个总体的身高分布是相同的。计算K-W 检验统计量并于卡方统计量比较（自由度k-1 = 3）：$KW=\frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}-3(n+1)$，纠正系数$C=1-\frac{\sum\left(\tau_j^3-\tau_j\right)}{n^3-n}$，$KW_c=\frac{KW}{C}$

1. 非参数Kruskal-Wallis检验

- kruskal-wallis函数，用来做Kruskal-Wallis（KW）检验（单因素非参数方差分析），检验的原假设是：k个独立样本来自于相同的总体。调用格式如下：
  - [p,table,stats]=kruskalwallis(X,group,displayopt)： X样本观测值矩阵，检验X的各列是否来自相同的总体； X是一个$m \times n$的矩阵， X的每一列是一个独立的样本，包含m个相互独立的观测。返回检验 p 值，如果 p 小于等于显著性水平，拒绝原假设，否则接受原假设。
  - 当X是一个矩阵时，用group参数（一个字符数组或字符串元胞数组）设定箱线图的标签，group的每一行（或每个元胞）与X的每一列对应，也就是说group的长度等于X的列数。如果X是一个向量，此时用group来指定X的每个元素（观测值）所在的组。
  - 当kruskalwallis函数给出的结果拒绝了原假设，则在后续的分析中，可以调用multcompare函数，把stats作为它的输入参数，进行多重比较。

1. 非参数Kruskal-Wallis检验

例1：某灯泡厂有四种不同配料方案制成的灯丝生产四批灯泡，每一批中随机抽取若干个做寿命试验，寿命数据如下表：

灯丝配料方案	灯泡寿命/h
A1	1600 1610 1650 1680 1700 1720 1800
A2	1580 1640 1600 1650 1660
A3	1460 1550 1600 1620 1640 1610 1540 1620
A4	1510 1520 1530 1570 1600 1680

根据表中数据分析灯丝的不同配料方案对灯泡寿命有无显著影响。显著性水平为0.05。

灯泡寿命通常不服从正态分布，不满足参数方差分析的基本假定，应该做非参数检验，下面调用kruskalwallis函数作非参数KW检验，调用anova1函数作参数检验，对比检验结果。

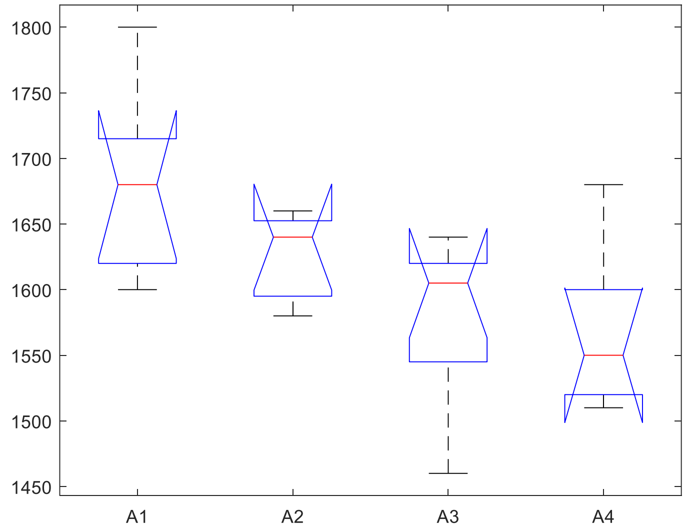
1. 非参数Kruskal-Wallis检验

检验的原假设：灯丝的不同配料方案对灯泡寿命无显著影响。

```
>> A1=[1600,1610,1650,1680,1700,1720,1800]';  %第1种配料方案的灯泡的寿命,需要转置
>> g1=repmat({'A1'},size(A1));
>> A2=[1580,1640,1600,1650,1660]'; %第2种配料方案
>> g2=repmat({'A2'},size(A2));
>> A3=[1460,1550,1600,1620,1640,1610,1540,1620]';   %第3种配料方案
>> g3=repmat({'A3'},size(A3));
>> A4=[1510,1520,1530,1570,1600,1680]'; %第4种配料方案
>> g4=repmat({'A4'},size(A4));
>> life=[A1;A2;A3;A4]; %将4种配料方案的灯泡寿命放在一起构成一个向量
>> group=[g1;g2;g3;g4];
>> [p1,~,stats1]=kruskalwallis(life,group) %调用kruskalwallis函数作Kruskal-Wallis检验
```

Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Groups	564.791	3	188.264	9.7	0.0213
Error	890.209	22	40.464		
Total	1455	25			



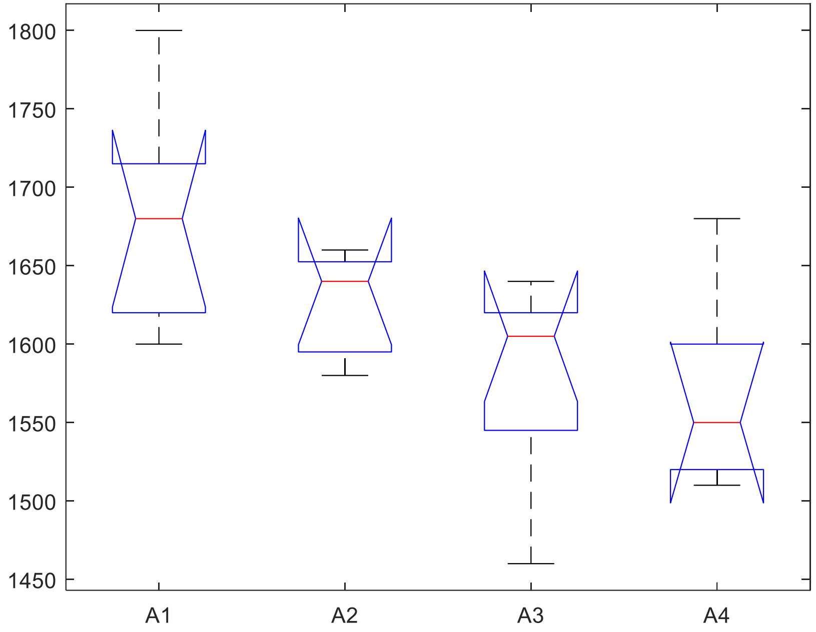
1. 非参数Kruskal-Wallis检验

%调用anova1函数作单因素一元方差分析

```
>> [p2,table2]=anova1(life,group)
p2 =
    0.0092
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	52950.5	3	17650.2	4.92	0.0092
Error	79003.3	22	3591.1		
Total	131953.8	25			



kruskalwallis函数返回的检验值p=0.0213<0.05，anova1函数返回的p值p=0.0092<0.05，说明在显著性水平0.05下，两种检验均拒绝了原假设，认为灯丝的不同配料方案对灯泡有显著影响。

1．非参数Kruskal-Wallis检验

为了进一步分析anova1函数和kruskalwallis函数的区别，即分析参数检验和非参数检验的区别，将A1方案中的1800改为2800，其他数据不变，然后再次调用这两个函数进行单因素一元方差分析。

```
%调用kruskalwallis函数作KW检验
>>[p1,table1,stats1]=kruskalwallis(life,group)
p1 =
    0.0213
%调用anova1函数作单因素一元方差分析
>> [p2,table2]=anova1(life,group)
p2 =
    0.1738
```

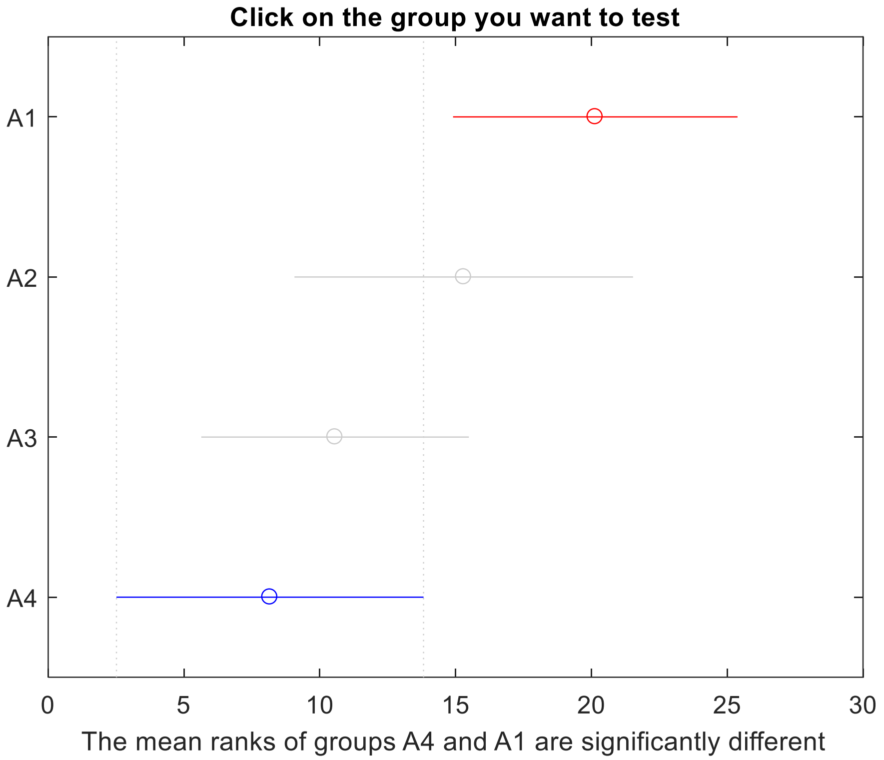
KW检验是基于秩的非参数检验，将样本观测数据中的最大值进一步增大，并没有改变样本的秩，所以两次调用kruskalwallis函数得到的结果完全相同，这说明KW检验不受个别异常值的影响。

而改变一个数据后调用anova1函数得到的结果与改变前是相反的，这反映了参数检验的局限性，当样本数据不满足参数方差分析的基本假定时，最好用非参数方差比较检验。

1. 非参数Kruskal-Wallis检验

多重比较：由于KW非参数检验认为灯丝是不同配料方案对灯泡寿命有显著影响，下面通过多重比较来检验在哪种配料方案下灯泡寿命的差异是显著的。

```
>> [c,m,h,gnames]=multcompare(stats1)
c =
    1.0000    2.0000   -6.6331    4.8429   16.3188    0.6993
    1.0000    3.0000   -0.5630    9.5804   19.7237    0.0722
    1.0000    4.0000    1.0724   11.9762   22.8800    0.0247
    2.0000    3.0000   -6.4356    4.7375   15.9106    0.6961
    2.0000    4.0000   -4.7344    7.1333   19.0010    0.4110
    3.0000    4.0000   -8.1888    2.3958   12.9804    0.9377
m =
   20.1429    2.8835
   15.3000    3.4117
   10.5625    2.6972
    8.1667    3.1145
```



从上面结果中可以看出，在显著性水平0.05下，灯丝的第1、4两种配料方案所对应的灯泡寿命的差异是显著的，其余配料方案所对应的灯泡寿命的差异是不显著的，并且第1种方案的平均秩最大，即灯丝的第一种配料方案所对应的灯泡的寿命最长。

2. 非参数Frideman检验

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例如：一项关于销售茶叶的研究报告说明三种销售方式可能和售出率有关。对一组商店在一段时间的调查结果列表如下（单位：购买者人数）。试问三种不同的销售方式是否有显著差异?

销售方式	购买率（%）							
商店内等待	20	25	29	18	17	22	18	20
门口销售	26	23	15	30	26	32	28	27
表演炒制	53	47	48	43	52	57	49	56

销售方式	对各项进行由低到高评秩（购买率（%））								合计
商店内等待	1	2	2	1	1	1	1	1	10
门口销售	2	1	1	2	2	2	2	2	14
表演炒制	3	3	3	3	3	3	3	3	24

$$\chi_r^2=\frac{12}{nk(k+1)}\sum R_j^2-3n(k+1)=\frac{12}{8\times3(3+1)}\left(10^2+14^2+24^2\right)-3\times8(3+1)=13>\chi_{0.05}^2(2)=5.99$$

2. 非参数Frideman检验

- friedman函数，用来做非参数Friedman检验（双因素方差分析）。调用格式：
  - [p,table,stats]=friedman(X,reps, displayopt)：根据样本观测值矩阵*X*进行均衡实验的非参数Fiedman检验。*X*的每一列对应参数A的一个水平，每行对应因素B的一个水平。reps表示因素A和B的每一个水平组合下重复的实验次数，默认值为1。
  - friedman函数检验矩阵*X*的各列是否来自于相同的总体，即检验因素A的各水平之间无显著差异，他对分组因素B不感兴趣。
  - Frideman函数返回检验的p值，当检验的p值小于或等于给定的显著性水平时，应拒绝原假设，原假设认为*X*总体来自于相同的总体。
  - frideman函数还生成1个图像，用来显示一个方差分析表。
  - 当friedman函数给出的结果拒绝了原假设，则在后续的分析中，可以调用multcompare函数，把stats作为它的输入，进行多重比较。

2. 非参数Frideman检验

例2：设有来自A、B、C、D四个地区的四名厨师制作名菜：水煮鱼，想比较他们的品质是否相同。四位美食评论对四名厨师的菜品分布做出了评分，如下表：

美食评委	地区			
	A	B	C	D
1	85	82	82	79
2	87	75	86	82
3	90	81	80	76
4	80	75	81	75

根据表中的数据检验四个地区制作的水煮鱼这道菜的品质有无差别，显著性水平为0.05

检验的原假设：四个地区的水煮鱼这道菜的品质没有区别。

2. 非参数Frideman检验

```
>> X = [85 82 82 79;87 75 86 82;90 81 80 76;80 75 81 75];
%调用friedman函数作Frideman检验，返回检验的p值，方差分析表table和结构体变量stats
>> [p,~,stats]=friedman(X)
p =
    0.0434
stats =
      source: 'friedman'
           n: 4
   meanranks: [3.7500 2 2.8750 1.3750]
       sigma: 1.2583
```

Friedman's ANOVA Table

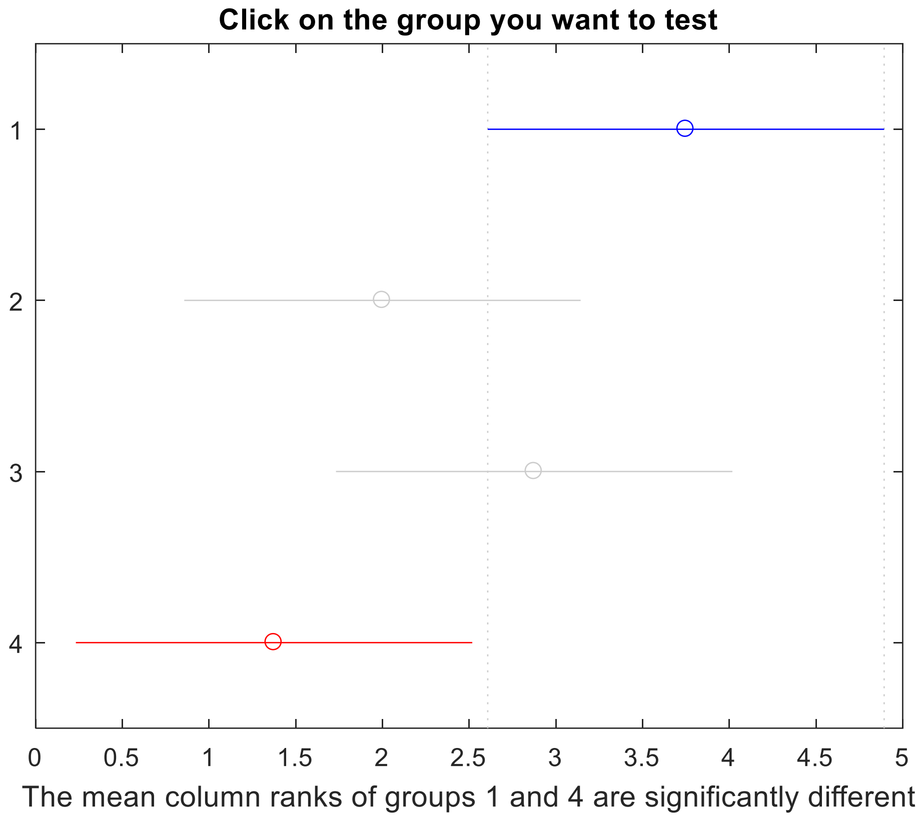
Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	12.875	3	4.29167	8.13	0.0434
Error	6.125	9	0.68056		
Total	19	15			

Test for column effects after row effects are removed

返回的检验p值p=0.0434<0.05，说明在显著性水平0.05下拒绝原假设，认为四个地区制作的水煮鱼的品质有显著性差别。具体是哪两个地区制作的水煮鱼这道菜的品质有显著差别，还需要作多重比较。

2. 非参数Frideman检验

```
>> [c,m,h,gnames]=multcompare(stats)
c =
    1.0000    2.0000   -0.5358    1.7500    4.0358    0.2006
    1.0000    3.0000   -1.4108    0.8750    3.1608    0.7589
    1.0000    4.0000    0.0892    2.3750    4.6608    0.0381
    2.0000    3.0000   -3.1608   -0.8750    1.4108    0.7589
    2.0000    4.0000   -1.6608    0.6250    2.9108    0.8962
    3.0000    4.0000   -0.7858    1.5000    3.7858    0.3311
m =
    3.7500    0.6292
    2.0000    0.6292
    2.8750    0.6292
    1.3750    0.6292
```



从以上结果可以看出，c矩阵的第3行的第3列和第5列构成的区间不包括0，说明在显著性水平0.05下，可认为A，D两个地区制作的水煮鱼这道菜的品质之间的差异是显著的。


01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听