北京建筑大学

理学院　　信息与计算科学专业　　实验报告

课程名称　《数据分析》　实验名称　方差分析与非参数检验　日 期　2020.4.24

班级　信171　姓名　李金哲　学号　201707010119　指导教师　王恒友　成绩

【实验目的】

（1）熟悉数据的基本统计与非参数检验分析方法；

（2）熟悉撰写数据分析报告的方法；

（3）熟悉常用的数据分析软件SPSS。

【实验要求】

根据各个题目的具体要求，完成实验报告。

【实验任务及结果与分析】

1、附件给出某年房屋价格的相关数据，请选用恰当的分析方法，对影响房屋价格的因素进行分析。（注意数据要调整成标准的格式，变量值、组别（字符变量转换成数值变量））（单因素方差分析选择其中两个因素分别进行分析、双因素方差分析选择其中任一对因素即可）

- **实验操作**

(1)选择菜单：“分析”->“比较均值”->“单因素 ANOVA”；

(2)选择观测变量“均价”到“因变量列表”框中；

(3)选择控制变量“环线位置”和“装修状况”先后添加到“因子”框中。

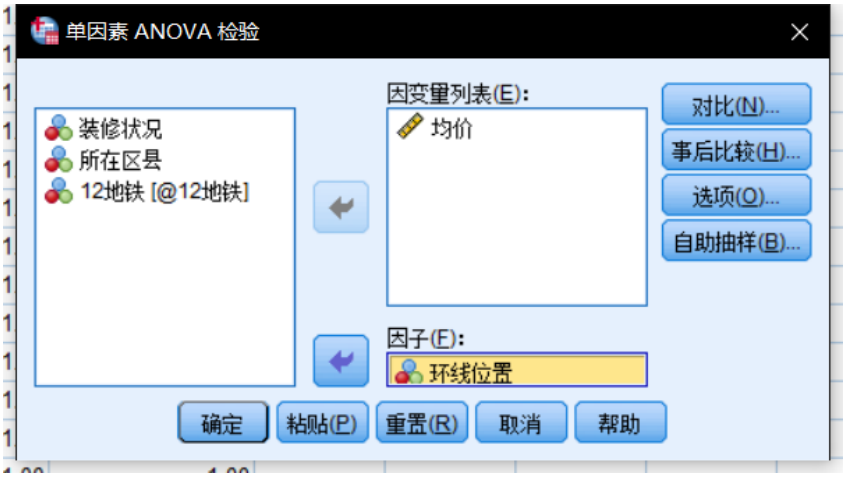

图1.1 “单因素方差分析”操作界面

分析结果如表1-1(a)和1-1(b)所示。

表1-1(a) 环线位置对均价影响的单因素方差分析结果

ANOVA					
均价					
	平方和	自由度	均方	F	显著性
组间	112.120	4	28.030	25.344	.000
组内	197.974	179	1.106		
总计	310.094	183			

表1-1(b) 装修状况对均价影响的单因素方差分析结果

ANOVA					
均价					
	平方和	自由度	均方	F	显著性
组间	79.180	1	79.180	62.408	.000
组内	230.914	182	1.269		
总计	310.094	183			

表1-1(a)是环线位置对均价影响单因素方差分析结果。可以看到：如果仅考虑环线位置单个因素的影响，则均价总变差310.094中环线位置可解释的变差为112.120，抽样误差引起的变差为197.974，它们的方差分别为28.030和1.106，相除所得的F统计量的观测值为25.34，对应的概率P值近似为0。**如果显著性水平α为0.05，由于概率P值小于显著性水平α，应拒绝原假设，认为环线位置对均价的平均值产生了显著影响，不同环线位置对均价的影响效应不全为0。**

表1-1(b)是装修状况对均价影响的单因素方差分析结果。可以看到：观测变量均价的离差平方总和为310.094；如果仅考虑装修状况单个因素的影响，则均价总变差中，装修状况可解释的变差为79.180，抽样误差引起的变差为230.914，它们的方差分别为79.180和1.269，相除所得的F统计量的观测值为62.408，对应的概率P值近似为0。**如果显著性水平α为0.05，由于概率P值小于显著性水平α，应拒绝原假设，认为装修状况对均价的平均值产生了显著影响，不同装修状况对均价的影响效应不全为0。**

对以上两个单因素进行进一步分析：

- **具体操作：**

(1)在图1.1所示窗口中点击“选项”按钮，结果如图1.2所示。

(2)在图1.2所示的窗口中，选择“描述性”、“方差同质性检验”和“平均值图”选项，“缺失值”选择“按分析顺序排除个案”。

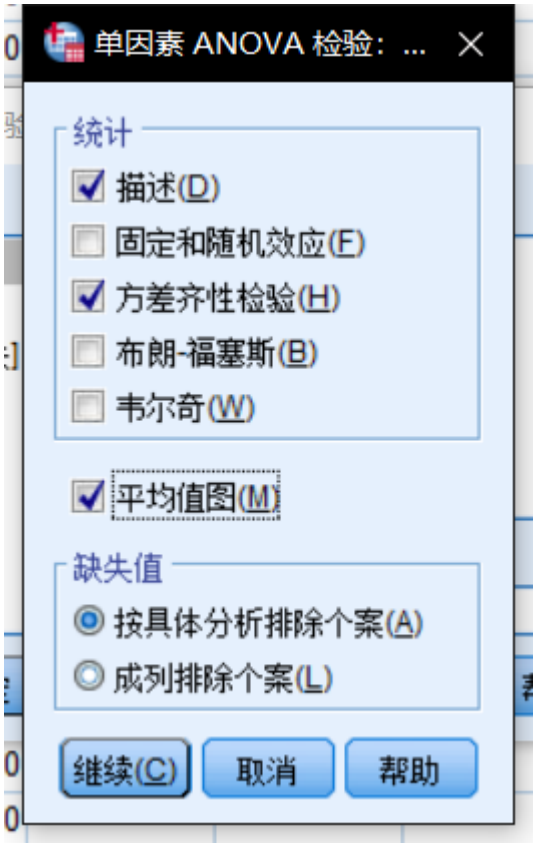

图1.2 “单因素方差分析：选项”操作界面

分析结果如表1-2(a)、1-2(b)、1-2(c)和1-2(d)所示。

表1-2(a) 环线位置对均价的基本描述统计量及95%置信区间

描述						
均价						

	个案数	平均值	标准差	标准误差	平均值的 95% 置信区间		最小值	最大值
					下限	上限		
1.00	26	4.338461538000001	.992401904000000	.194626026000000	3.937621735000001	4.739301342000000	2.500000000000000	6.300000000000001
2.00	60	3.686833333000000	1.549060395000000	.199982837000000	3.286668601000000	4.086998066000001	1.800000000000000	8.600000000000001
3.00	30	3.194333333000000	.663842957000000	.121200587000000	2.946450299000000	3.442216367000000	2.200000000000000	4.400000000000000
4.00	60	2.416833333000000	.547321650000000	.070658921200000	2.275445159000000	2.558221508000000	1.600000000000000	3.900000000000000
5.00	8	1.347500000000000	.434634493000000	.153666499000000	.984136470000000	1.710863530000000	.800000000000000	1.800000000000000
总计	184	3.182771739000000	1.301730430000000	.095964840500000	2.993431965000000	3.372111513000000	.800000000000000	8.600000000000001

表1-2(a)中，“1”“2”“3”“4”“5”分别对应环线“2至3环”“3至4环”“4至5环”“5至6环”“6环以外”在5个环线中各有26、60、30、60、8个样本。2至3环的均价最高，3至4环与4至5环居中，5至6环其次，6环以外最低。这些结论同样可在图1.3中印证。

表1-2(b) 不同环线位置的方差齐性检验结果

方差齐性检验			
均价			
莱文统计	自由度 1	自由度 2	显著性
7.970	4	179	.000

表1-2(b)表明，如果显著性水平α为0.05，由于概率P值小于显著性水平α，**因此应拒绝原假设，认为不同环线下对均价的总体方差有显著差异。**

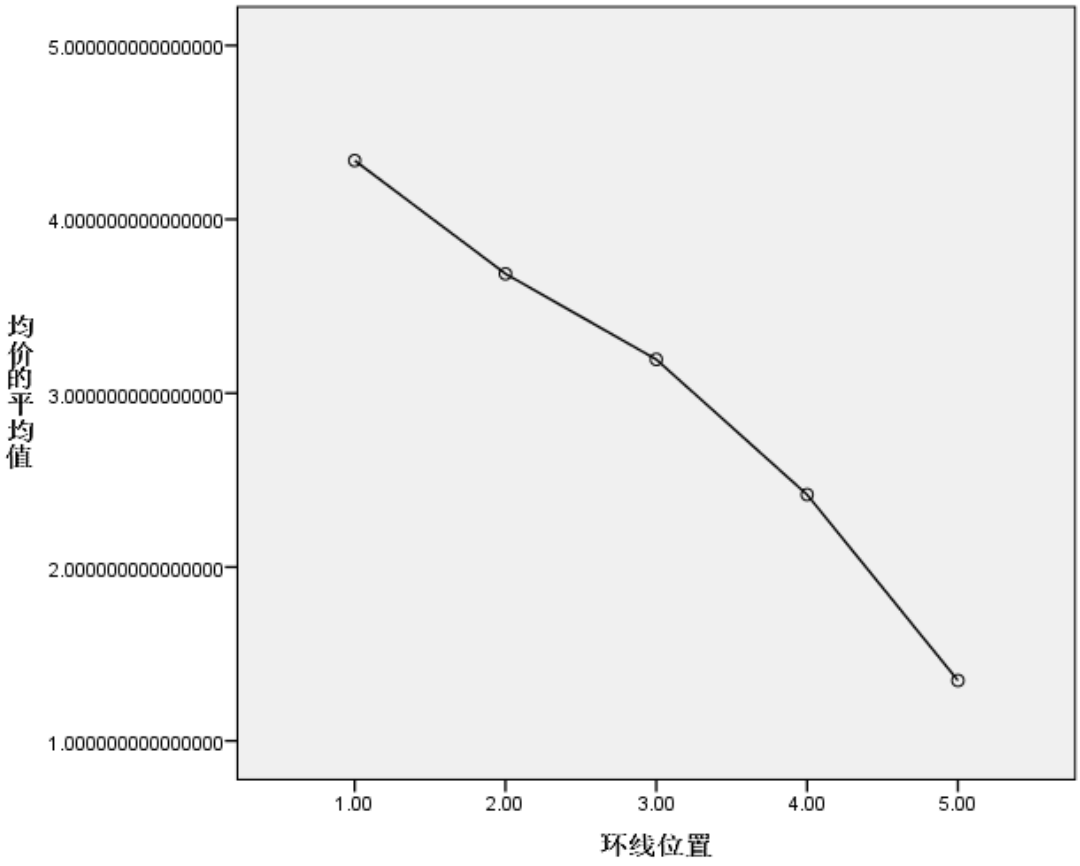

图1.3(a) 不同环线位置均价的均值折线图

表1-2(c) 装修状况对均价的基本描述统计量及95%置信区间

描述								
均价								
					平均值的 95% 置信区间			
	个案数	平均值	标准差	标准误差	下限	上限	最小值	最大值
1.00	84	2.46702381 0000000	.579667083 000000	.063246863 900000	2.34122834 1000000	2.59281927 8000000	.800000000 000000	3.90000000 0000000
2.00	100	3.78400000 0000000	1.43204452 3000000	.143204452 000000	3.49985129 8000000	4.06814870 2000000	1.00000000 0000000	8.60000000 0000001
总计	184	3.18277173 9000000	1.30173043 0000000	.095964840 500000	2.99343196 5000000	3.37211151 3000000	.800000000 000000	8.60000000 0000001

表1-2(c)表明，在2个装修状况下下分别有84、100两个样本。“2”即“精装修”的平均均价高于“1”即“毛胚”可在图1-3(b)中得到印证。

表1-2(d) 装修状况方差齐性检验结果

方差齐性检验			
均价			
莱文统计	自由度 1	自由度 2	显著性
28.807	1	182	.000

表1-2(d)表明,如果显著性水平α为0.05,由于概率P值大于显著性水平α,**因此应拒绝原假设，认为装修状况对均价的总体方差有显著差异。**

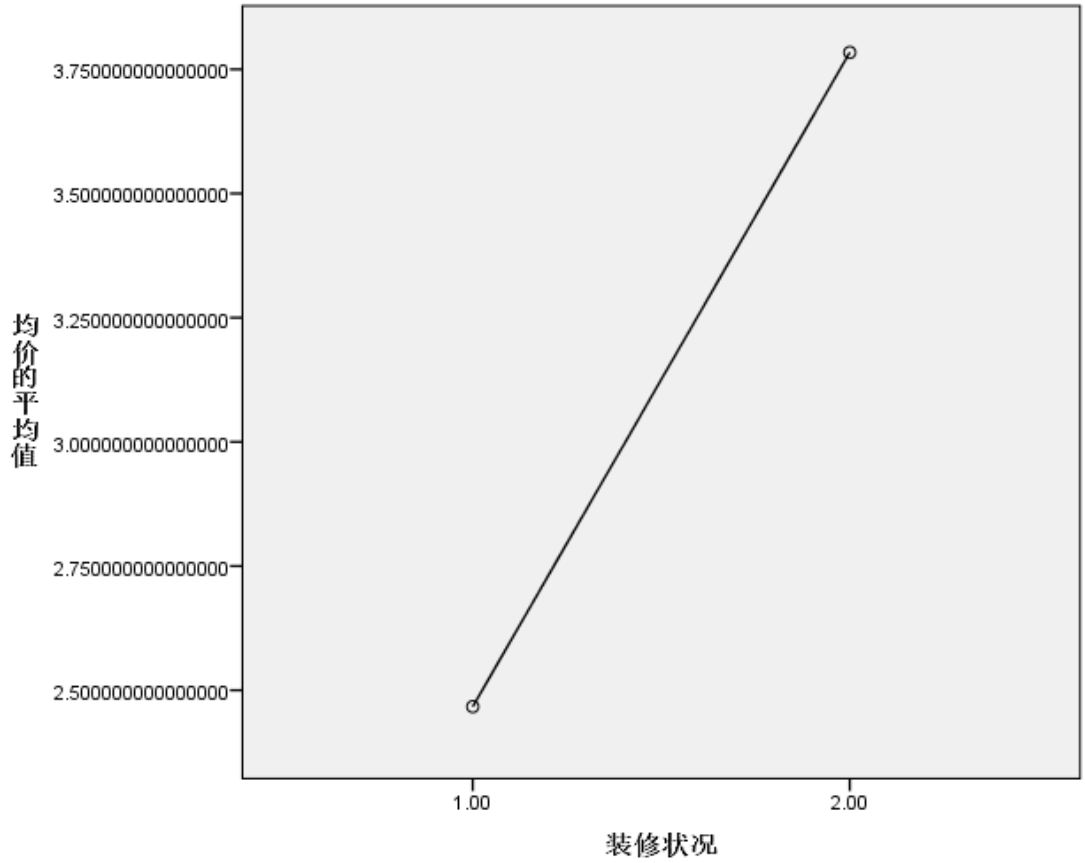

图1.3(b) 装修状况的均价的均值折线图

将以上两个单因素配对进行多因素分析：

- **具体操作：**
  （1）选择菜单：“分析”->“一般线性模型”->“单变量”；
  （2）将“均价”添加到“因变量”框中；
  （3）将“环线位置”和“装修状况”添加到“固定因子”框中。

图1.5 “多因素”操作界面

分析结果如表1-3所示。

表1-3 均价多因素方差分析结果

主体间效应检验					
因变量： 均价					
源	III 类平方和	自由度	均方	F	显著性
修正模型	146.478[a]	9	16.275	17.308	.000
截距	590.549	1	590.549	628.029	.000
环线位置	47.605	4	11.901	12.657	.000
装修状况	15.435	1	15.435	16.414	.000
环线位置 * 装修状况	6.849	4	1.712	1.821	.127
误差	163.616	174	.940		
总计	2174.021	184			
修正后总计	310.094	183			
a. R 方 = .472（调整后 R 方 = .445）					

表1-3中，可以看到：观测变量的总变差SST为310.094，它被分解为三个部分，分别是：由有环线位置引起的变差47.605，由装修状况引起的变差15.435，由有环线位置和装修状况交互作用引起的变差6.849，由随机因素引起的变差163.616。这些变差除以各自的自由度后，得到各自的方差，并可计算出各F检验统计量的观测值和一定自由度下的概率P值，均为0。如果显著性水平α为0.05，对于**环线位置**而言，由于其概率P值小于显著性水平α，所以**应拒绝原假设，可以认为不同环线位置的均价总体均值存在显著差异**，对于**装修状况**而言，由于其概率P值小于显著性水平α，所以**应拒绝原假设，可以认为装修状况的均价总体均值存在显著差异**。该结论与单因素方差分析是一致的。但由于环线位置和

装修状况交互作用的概率P值大于显著性水平α，因此**不应拒绝原假设，可以认为环线位置和装修状况没有对均价产生显著的交互作用，不同装修状况的房屋周边的环线位置都不会对均价产生显著影响。**

2、附件给出管理才能评分的相关数据，请选用恰当的分析方法，分析该评分数据是否服从正态分布。

- **具体操作：**

(1) 选择菜单：“分析”→“非参数检验”→“旧对话框”→“单样本K-S”；

(2) 选择“管理才能评分”到“检验变量列表”框中；

(3) 在“检验分布”选择“正态”。

图2.1 “单样本K-S”操作界面

分析结果如表2-1所示

表2-1 管理才能评分总体分布的K-S检验结果

单样本柯尔莫戈洛夫-斯米诺夫检验		
		管理才能评分
个案数		90
正态参数[a,b]	平均值	487.6778
	标准差	88.28005
最极端差值	绝对	.066
	正	.066
	负	-.041
检验统计		.066
渐近显著性（双尾）		.200[c,d]
a. 检验分布为正态分布。		
b. 根据数据计算。		
c. 里利氏显著性修正。		
d. 这是真显著性的下限。		

表2-1表明，数据的均值为487.6778，标准差为88.28005。最大绝对差值为0.066，最大正差为0.066，最小负差为-0.041，概率P值为0.200。如果显著性

水平α为0.05，由于其概率P值大于显著性水平α，所以**不应拒绝原假设，认为该评分数据的总体分布为正态分布的假设。**

3、附件给出了某体育比赛的两位裁判打分数据，请选用恰当的分析方法，检验该两组评分分布是否有显著差异。(注意数据要调整成标准的格式，变量值、组别)

- **具体操作：**

(1) 选择“分析”->“非参数检验”->“旧对话框”->“2个独立样本”来对数据进行分析。

(2) 选择“得分”到“检验变量列表”框中；

(3) 选择选择“组别”到“分组变量”框中并给定义组中的两个组标记“1”和“2”；

(4) 在“检验类型”中选择“曼-惠特尼U检验”。

图3.1 “U检验”操作界面

表3-1（a） 裁判打分的U检验结果

秩				
	组别	个案数	秩平均值	秩的总和
得分	中国	31	27.77	861.00
	美国	29	33.41	969.00
	总计	60		

表3-1(a)中，可以看到：从1、2两组中，即中美裁判中分别抽取了31和29个样本两个秩和分别为861和969；W统计量应采取美国裁判的秩和Wx；

表3-1（b） 裁判打分的U检验结果

检验统计[a]	
	得分
曼-惠特尼 U	365.000
威尔科克森 W	861.000
Z	-1.253
渐近显著性（双尾）	.210
a. 分组变量：组别	

表3-1(b)中，U，Z统计量分别为365和-1.253。由于是小样本，因此采用U统计量的精确概率。如果显著性水平α为0.05，由于其概率P-值大于显著性水平α，所以**不拒绝原假设，认为中美裁判打分不存在显著差异。**

4、附件给出了减肥茶数据，请选用恰当方法分析，检验该减肥茶是否对减肥有显著效果。(注意数据要调整成标准的格式，变量值、组别)

- **具体操作：**

(1) 选择“分析”->“非参数检验”->“旧对话框”->“2个相关样本”来对数据进行分析。

(2) 选择“喝茶前体重”和“喝茶后体重”到“检验对”框中；

(3) 在“检验类型”中选择“威尔科克森”。

图4.1 “两个关联样本检验”操作界面

分析结果如表4-1(a)、4-1(b)所示。

表4-1(a) 喝茶前后体重两配对样本威尔克特森检验结果

秩				
		个案数	秩平均值	秩的总和
喝后体重 - 喝茶前体重	负秩	44[a]	23.38	1028.50
	正秩	1[b]	6.50	6.50
	绑定值	0[c]		
	总计	45		
a. 喝后体重 < 喝茶前体重				
b. 喝后体重 > 喝茶前体重				
c. 喝后体重 = 喝茶前体重				

由表4-1(a)可知，负号秩总和为44，意味着喝茶后体重低于喝茶前体重的有44人，正号秩总和为1，表明体重远高于喝茶前的有1人。

表4-1(b) 喝茶前后体重两配对样本威尔克特森检验结果

检验统计[a]	
	喝后体重 - 喝茶前体重
Z	-5.771[b]
渐近显著性（双尾）	.000
a. 威尔科克森符号秩检验	
b. 基于正秩。	

由表4-1(b)可知，双侧的二项分布累计概率为0。如果显著性水平α为0.05，由于其概率P-值小于显著性水平α，所以**拒绝原假设，认为喝减肥茶前后的体重分布有显著差异，喝减肥茶有显著效果。**