# 第一阶段 业务分析师

# 建模数据分析师

## 网校相关视频

SPSS数据分析3

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* 方差分析
* 双因素方差分析
* 卡方检验
* Rxc卡方检验
* 时间序列分析
* 数据分析的思路

### 【重点】

* 方差分析
* 双因素方差分析
* 卡方检验
* Rxc卡方检验
* 时间序列分析
* 数据分析的思路

### 【难点】

* 方差分析
* 双因素方差分析
* 卡方检验
* Rxc卡方检验
* 时间序列分析

## 复习巩固作业讲解

无

## 本次课程任务讲解

### 【知识点1】方差分析

1. 方差分析基本思想

进行多分类的均值比较

t检验：关注二分类变量对连续变量的影响

方差分析：关注多分类变量对连续变量的影响

本质上来说，方差分析研究的是分类自变量对连续型因变量的影响

1. 单因素方差分析

前置条件：

1）有1个多分类的自变量和1个连续型的因变量

2）各组间的观测值相互独立

3）每组内的观测值不能有明显的异常值

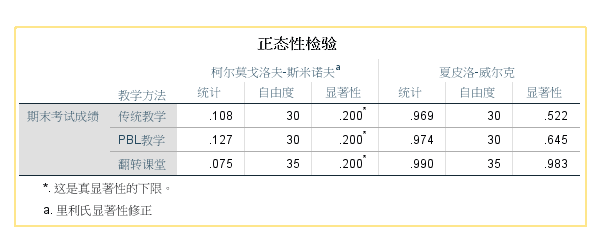
4）每组内的观测值要符合或者近似符合正态分布

5）每组的方差要相等

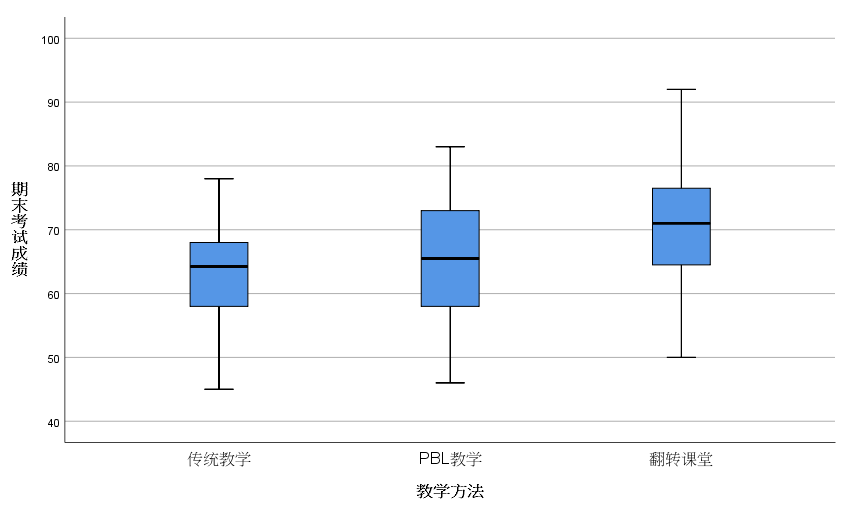
操作步骤：

数据源：单因素方差分析.sav

通过探索过程进行正态性/异常值检查

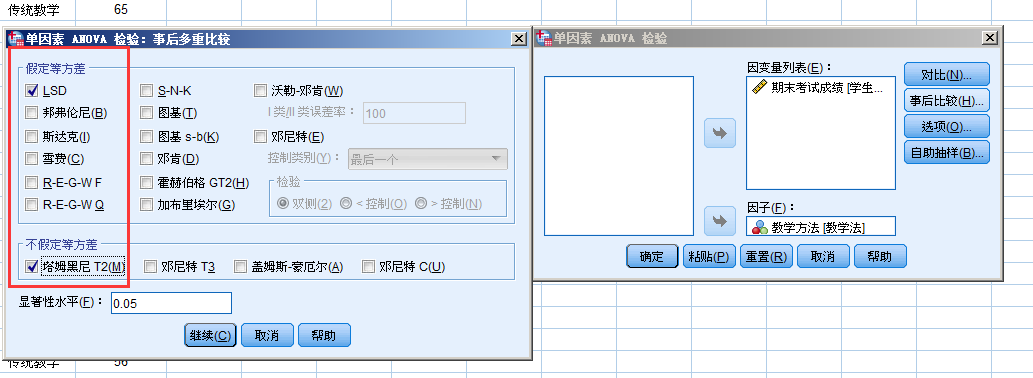


显著性都大于0.5，都符合正态分布



分析-比较均值-单因素ANOVA分析



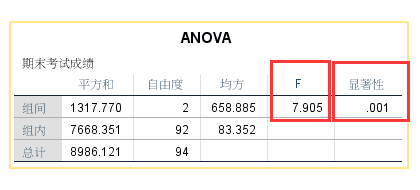


结果解读



方差齐性检验

显著性0.692>0.05,大同小异，意味着方差相等，通过了方差齐性检验



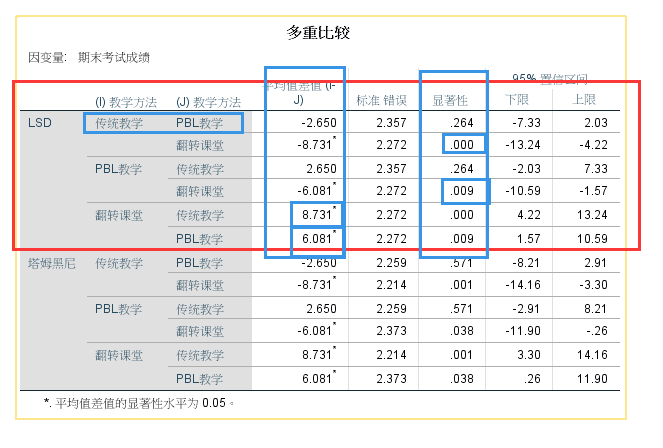
方差分析表，看F值和Sig(p)值

Sig<0.05,通过了方差检验，大同小异，三种教学方法存在显著差异

H0>0.05接受原假设，H1<=0.05,接受备择假设

原假设一般默认样本来自同一总体

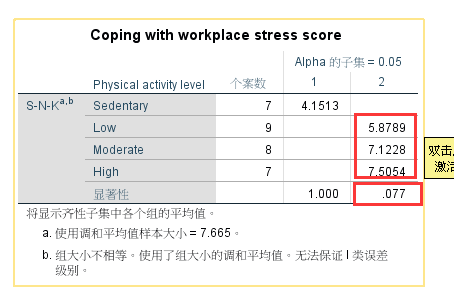
所以当前这个例子说明来自不同个体，三种教学模式成绩不同



组间比较表，方差齐看LSD方法，观察显著性明显时的平均值差值

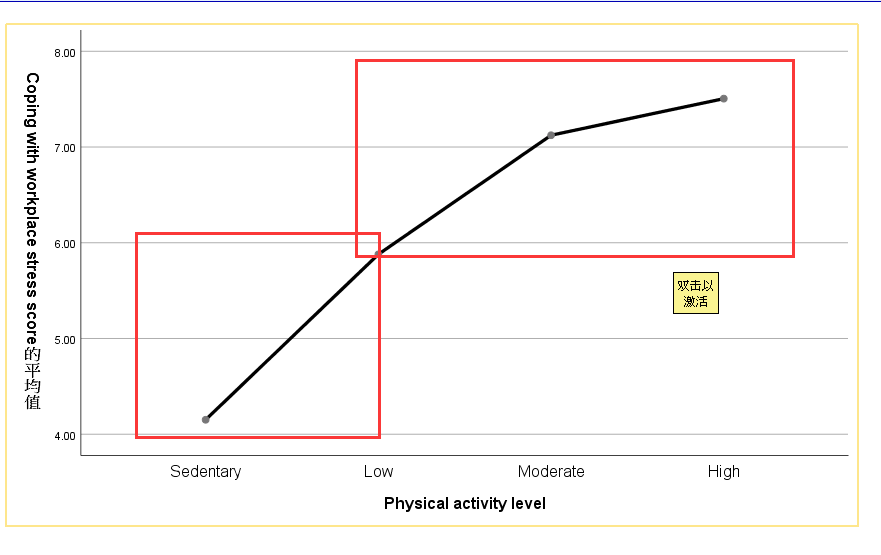
数据源：单因素方差分析2，探索行为和抗压能力的影响

亦可通过SNK或其他方法观察，或通过绘制平均值图进行可视化观察分析



1组：Swdentary久坐：4.15

2组：运动时间短中长：0.077>0.05，意味着运动时间长短中属于同一组别，没有显著性差异。也意味着久坐和运动时间中长短有显著性差异



久坐到运动时间短变化比较显著。

运动时间中长短变化没有特别显著

### 【知识点2】双因素方差分析

1.双因素方差分析

了解两个分类变量对1个连续变量的影响（两个分类变量之间的相关关系是否对连续变量会有影响）

前置条件

1）有两个多分类自变量，1个连续因变量

2）观测的独立性

3）任一分类中没有明显的异常值

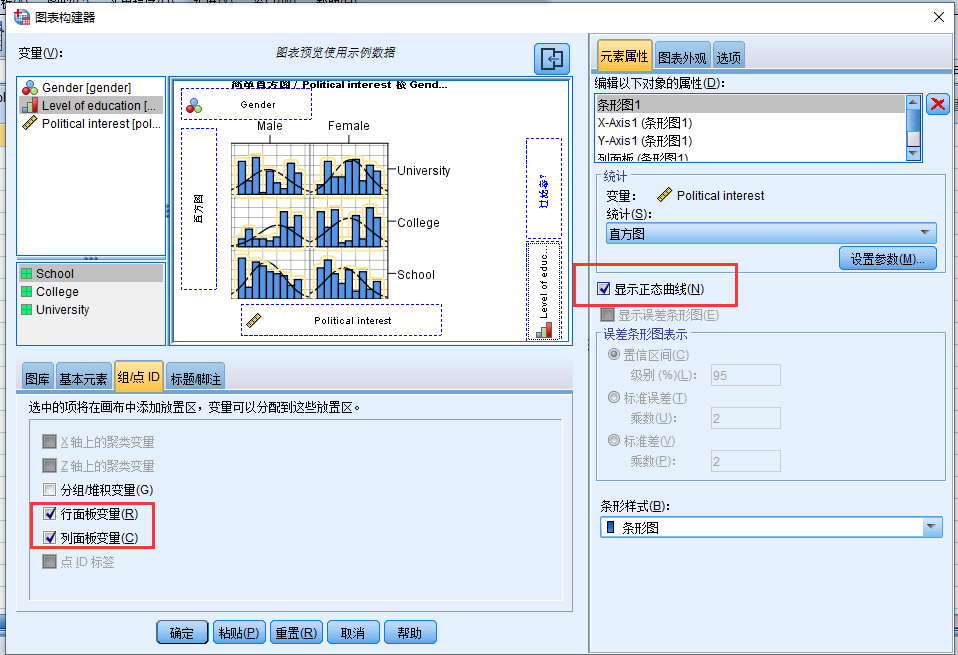
4）任一分类具有等方差性

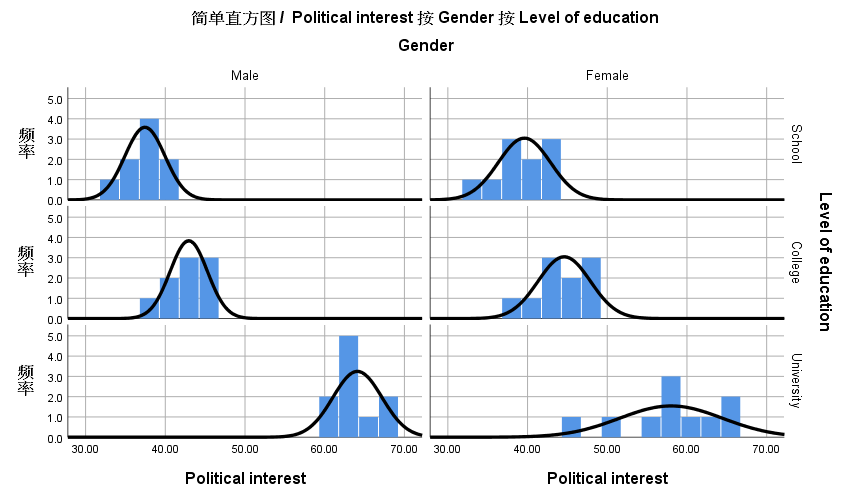
5）任一分类的观测值要近似符合正态分布

数据源：两因素方差分析，来分析不同性别不同教育程度对政治兴趣得分的影响

操作步骤

先通过直方图组观察/检测各组观测值分布情况





分析--一般线性模型--单变量（单变量是指因变量个数）



模型采用全因子



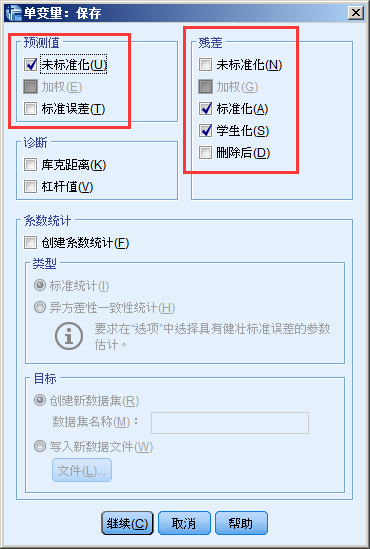
通过设定均值变化图观察交互作用的影响



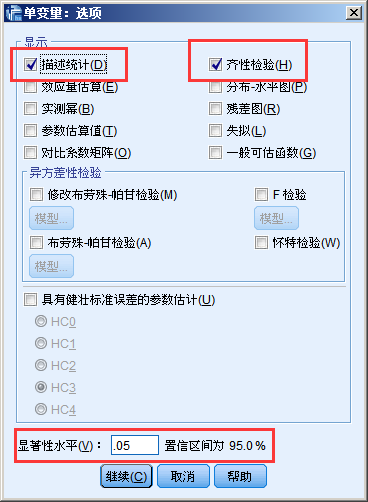
设定事后检验方法



如需预测或对残差进行检查，可以设置保存值

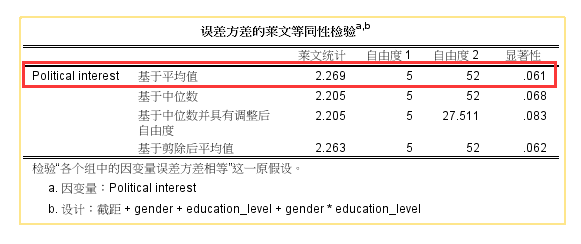


选项设置



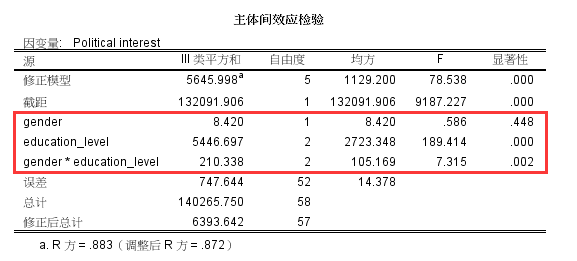
结果解读

方差齐性检验



显著性0.061<0.05，通过了显著性检验

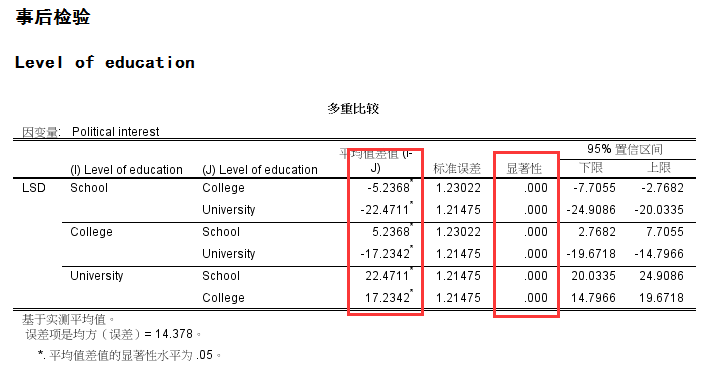
主体效应检验（相当于单因素方差分析的方差分析表）



性别显著性0.448>0.05，没有通过显著性检验，受教育程度和性别、受教育程度<0.05，通过了显著性检验。

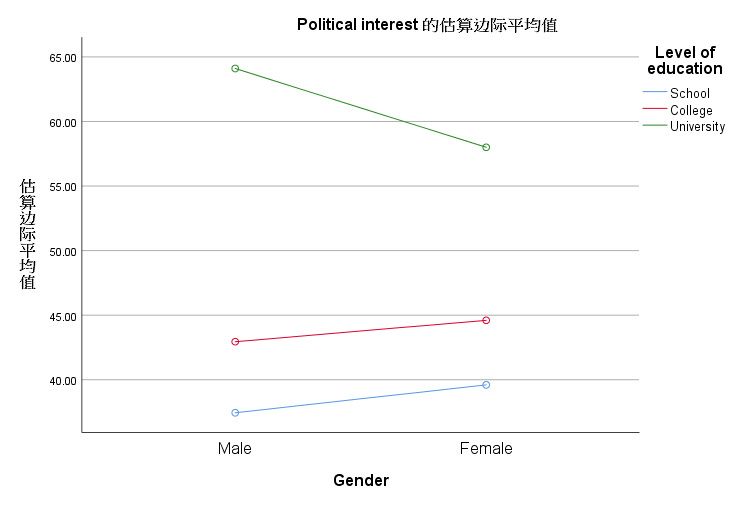
可见受教育程度和性别与受教育程度的交互项对政治兴趣有显著的影响（P<0.05）

事后检验可以观察多分类变量的组间差异

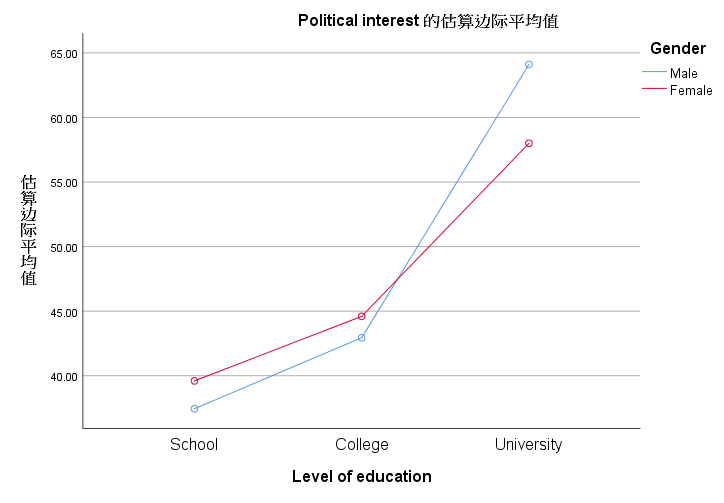


显著性都小于0.05，都有明显的差异性，说明不同的教育程度对政治兴趣得分有显著性的差异

也可以通过均值图进行可视化观察分析



大学的教育程度，男性的政治兴趣得分显著高于女性。



由图可见在提高受教育程度增加政治兴趣的过程中，男性比女性获益更大。

### 【知识点3】卡方检验

1. 卡方检验基本思想

研究分类变量对分类变量的影响

交叉表

行：分类变量

列：分类变量

交叉的单元格：频数

卡方值=（真实值-理想值）^2/理想值

1. 卡方检验的前置条件

有两个二分类变量（多分类变量）

观察值要相互独立

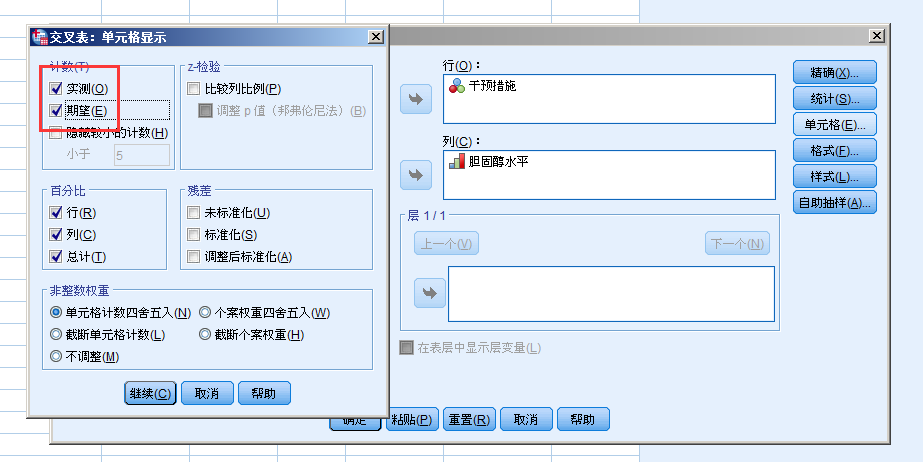
样本量要足够大，分析中任一个期望频数>5

1. 卡方检验基本操作

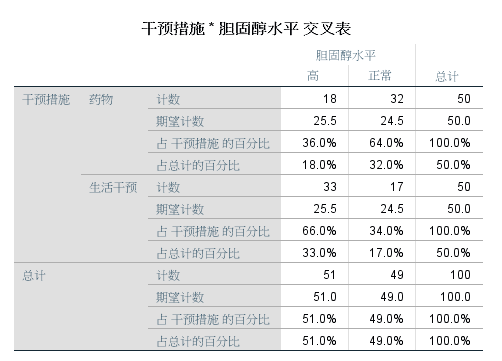
数据源：2x2卡方检验

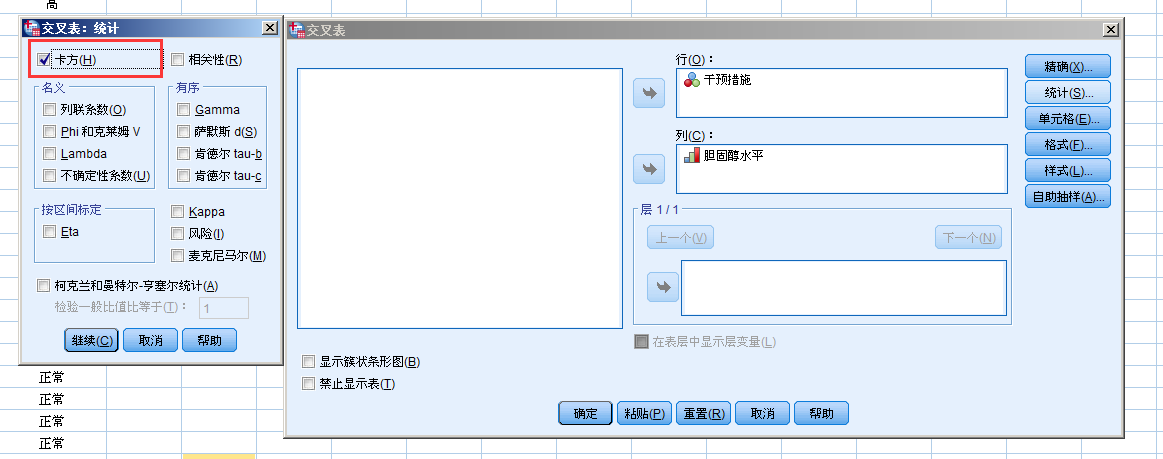
判断不同的干预措施对胆固醇水平的影响

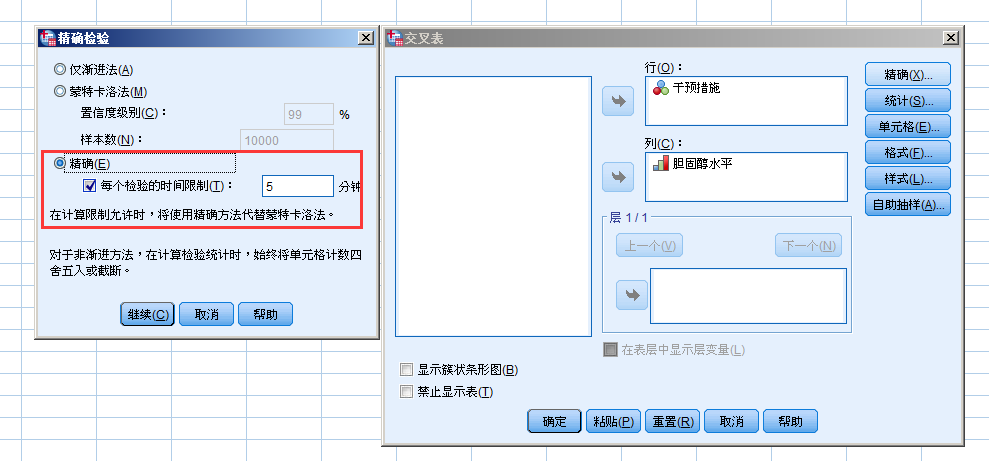
分析-描述统计-交叉表



观察期望值



观察卡方值



结果解读：



当a:0个单元格的期望小于5，并且有效个案数100>40的时候，看皮尔逊卡方值=9.004，看渐进显著性0.003<0.05，“大同小异”，通过了显著性检验，意味着两种干预措施对胆固醇治疗效果存在显著性差异。

注意：主要看：1、有效个案数N>40。2、期望值T>=5

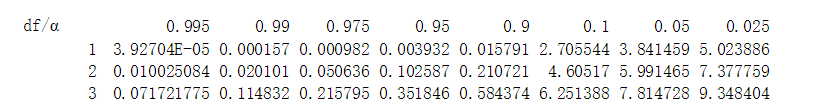
如果T<5,看费希尔精确检验

自由度=（行数-1）\*（列数-1）

例子1：3个数总和为5，可以任意填写2个数字，第三个数字固定，所以有n-1个自由度

例子2：

卡方检验临界值表：<https://wenku.baidu.com/view/26879a52a7c30c22590102020740be1e650eccbc.html>



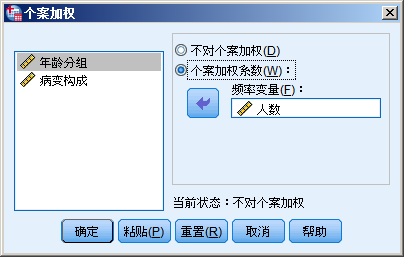
自由度为1，临界值为0.003932，卡方值>临界值，则实际观测值与理论推断值偏离程度越大。

卡方检验就是[统计样本](https://www.baidu.com/s?wd=%E7%BB%9F%E8%AE%A1%E6%A0%B7%E6%9C%AC&tn=SE_PcZhidaonwhc_ngpagmjz&rsv_dl=gh_pc_zhidao)的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，卡方值越大，越不符合；卡方值越小，偏差越小，越趋于符合，若两个值完全相等时，卡方值就为0，表明理论值完全符合。

### 【知识点4】Rxc卡方检验

注意针对有频数的交叉表，先用频数加权之后再做卡方检验（数据-个案加权）





前置条件：

1、有两个多分类变量

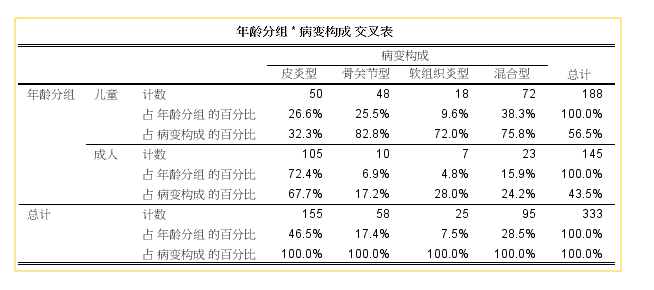
2、观察值要相互独立

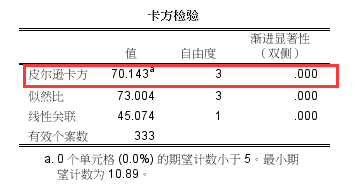
3、样本量要足够大，分析中任一个期望频数>5

基本操作：分析-描述统计-交叉表（和2X2卡方检验类似）

结果解读：

基本情况





N=333>40,看皮尔逊卡方，显著性=0.000<0.05,说明不同年龄分组中病变构成存在显著性差异

### 【知识点5】时间序列分析

思路

针对某一段时间中某个变量获得的一组观测值进行分析，往往是针对只有一个连续性的因变量（按照固定的时间间隔进行排列）而没有自变量的情况

1. 时间序列的预处理

数据源：时间序列1

针对缺失值进行填补（可采用多种方法，效果各有差异）

转换--替换缺失值

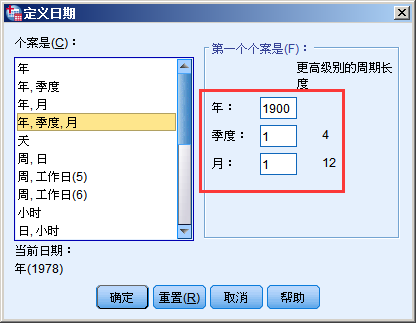


线性插值：相邻两个值的平均值

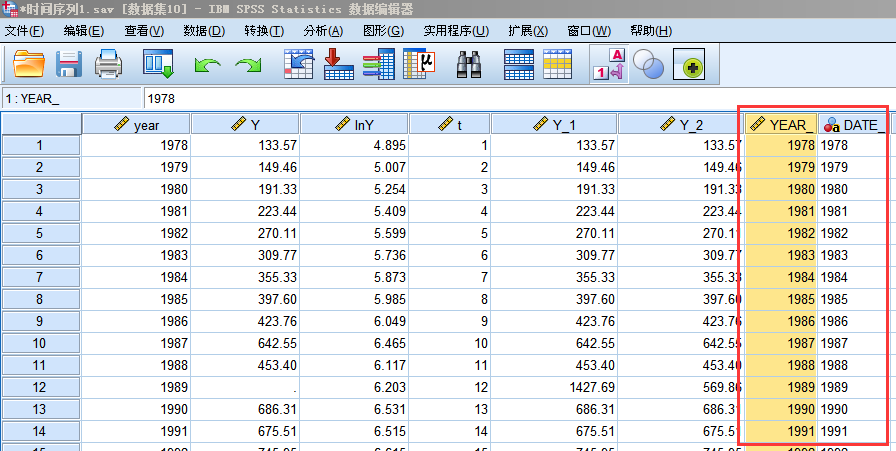
1. 创建时间序列变量

只有定义了时间变量，SPSS系统才能识别指定序列的时间变量，从而进行后续的时间序列分析

数据--定义日期和时间



设定日期变量的起始值。注意该操作会生成新的变量

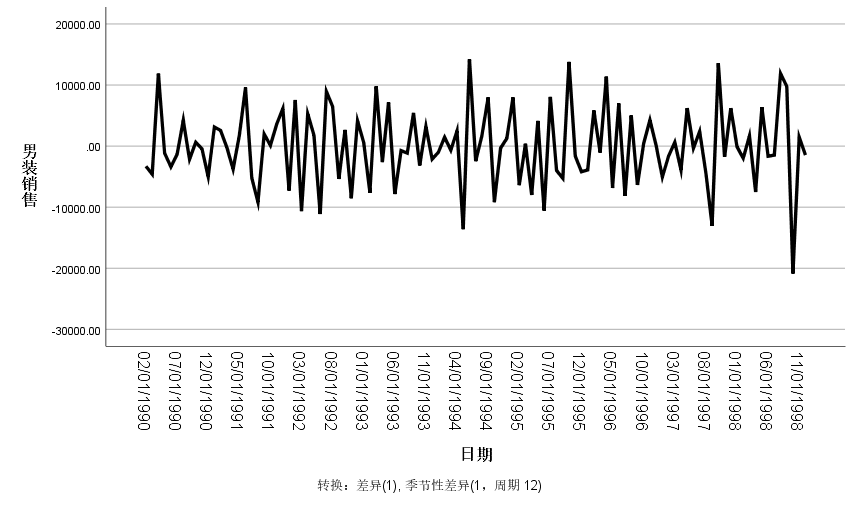


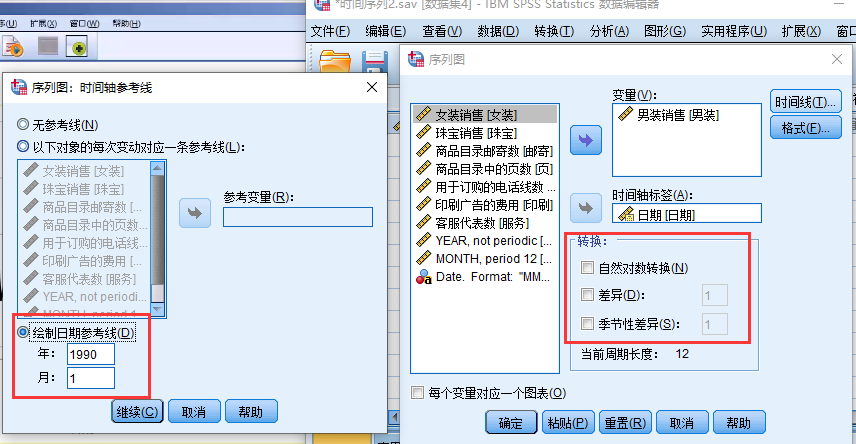
1. 时间序列图

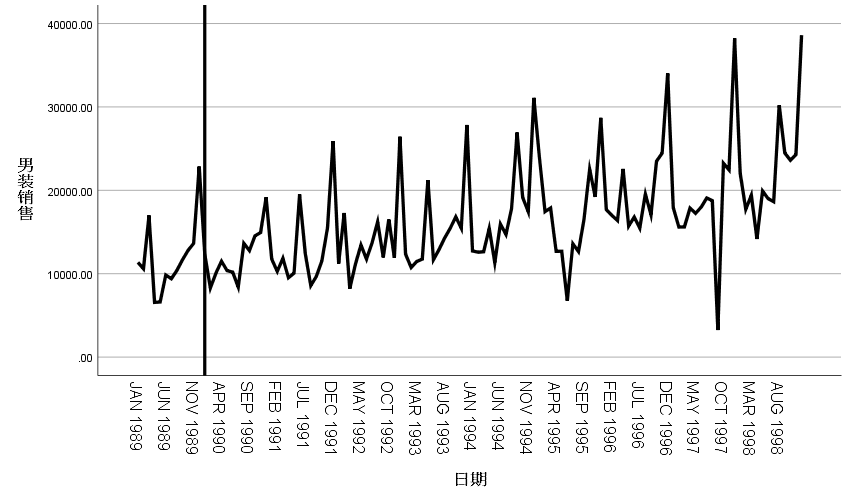
数据源：时间序列2，观察男装销售情况按照日期的变化趋势

作用：观察时间序列的变化情况以及去除时间序列中的变异情况后的平滑序列

分析-时间序列预测-序列图

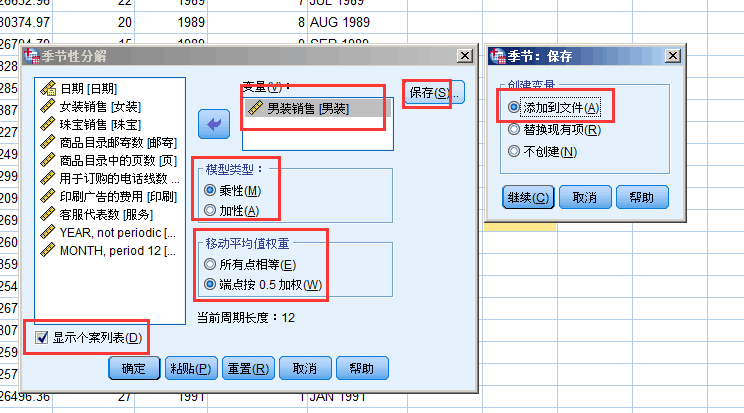






1. 季节性分解

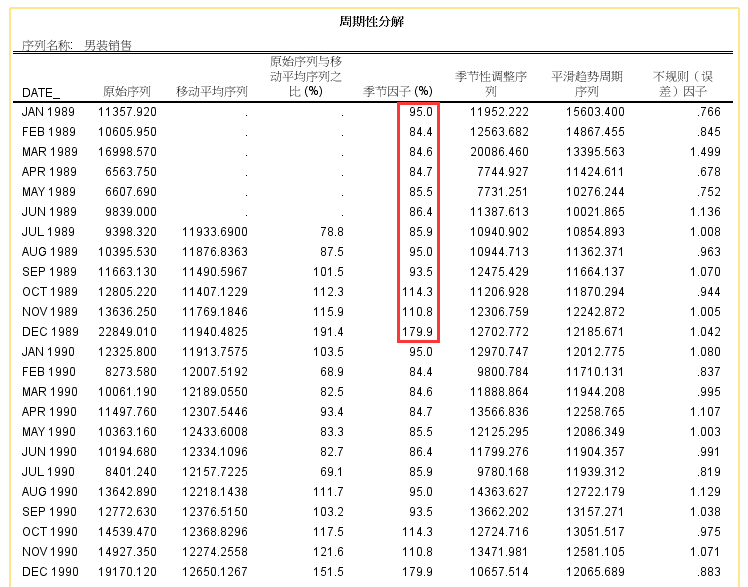
去除时间序列中的季节性变异情况，同时可以方便观察时间序列数据中的变化趋势



模型类型：变异添加到模型的方法，大多数情况默认为乘法

所有点相等：周期长度为奇数

端点按0.5加权：周期长度为偶数：月份为12个月



注意观察季节因子，可以发现男装销售呈现明显的季节趋势，12月销量最高

“季节性分解”过程创建了四个新变量（序列），并且每个指定的序列都带有以下三个字母的前缀：

SAF。季节性调整因子。这些值指示每个周期对序列水平的影响。

SAS。季节性调整序列。这些值是在删除序列的季节性变化之后获得的。

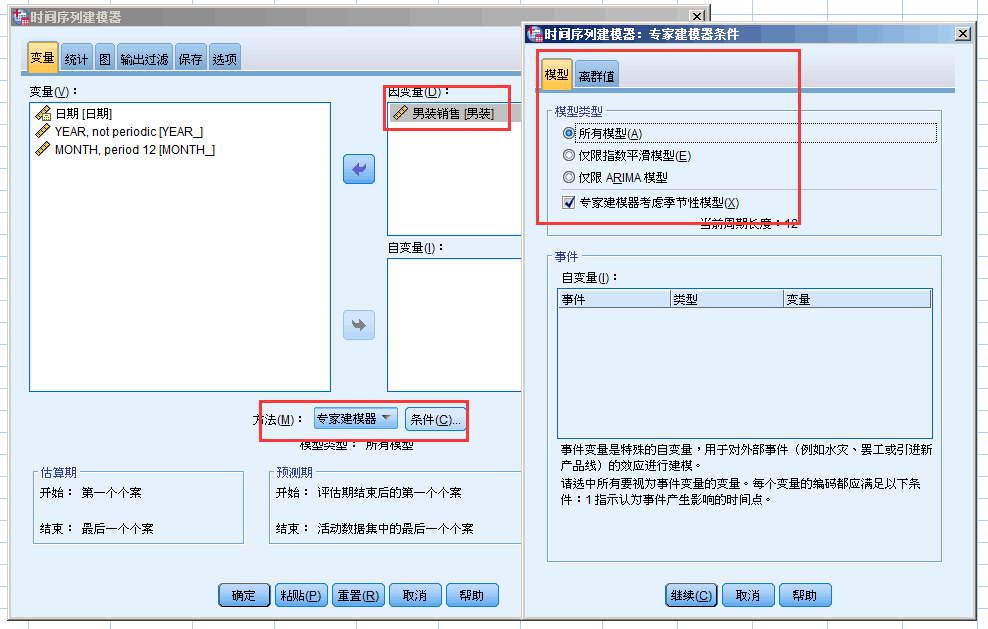
STC。平滑的趋势循环成分。这些值显示序列中出现的趋势和循环行为。

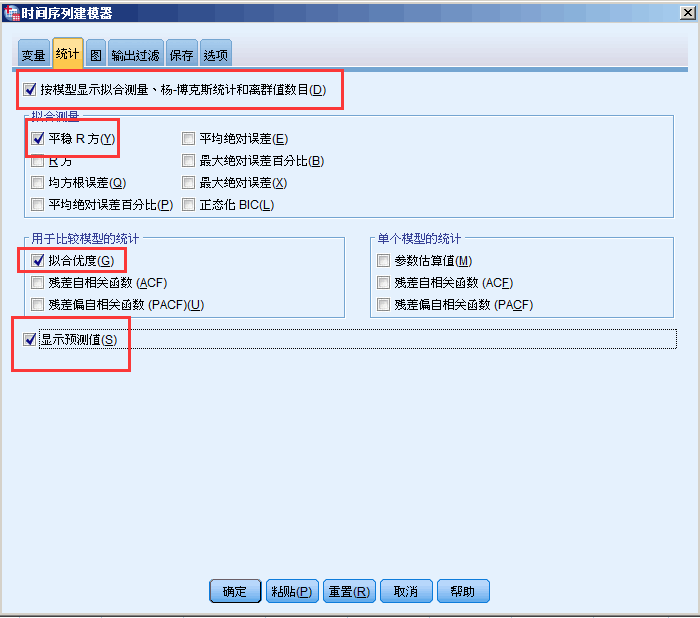
ERR。残差或“误差”值。这些值是在从序列中删除季节性、趋势和循环成分之后保留的。

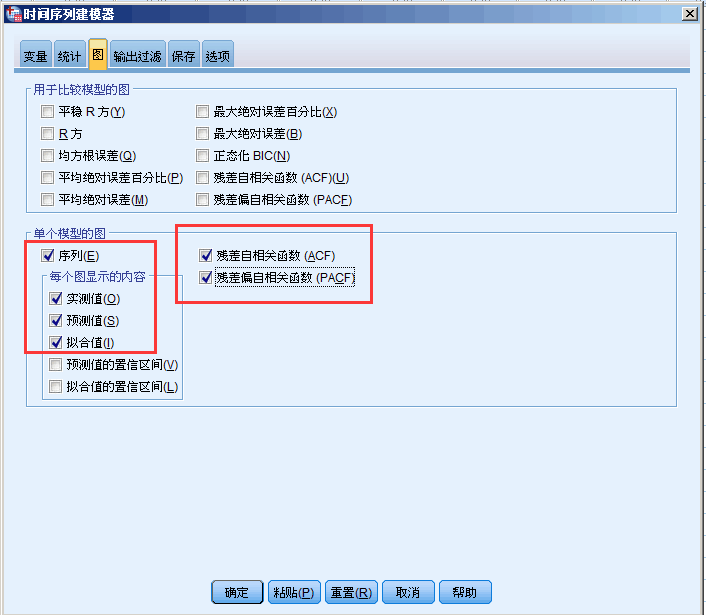
1. 创建时间序列模型

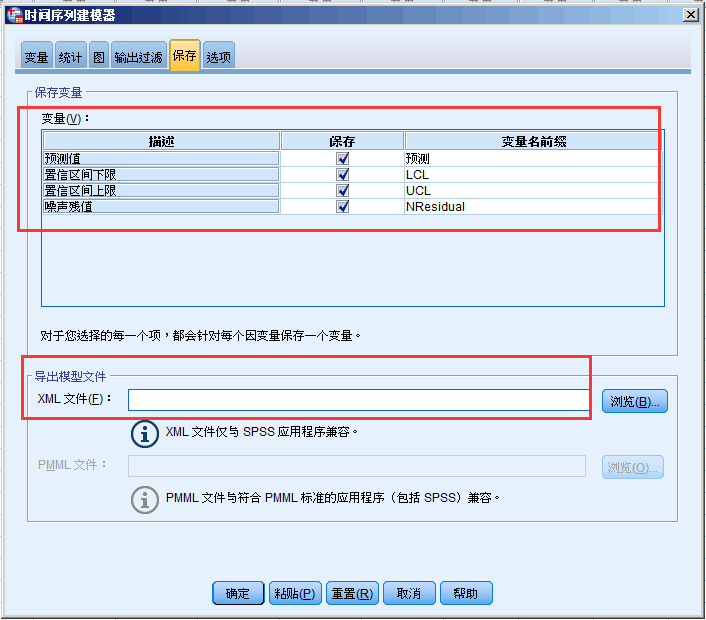
用于创建针对时间序列数据的预测模型，以便进行后续序列值的预测

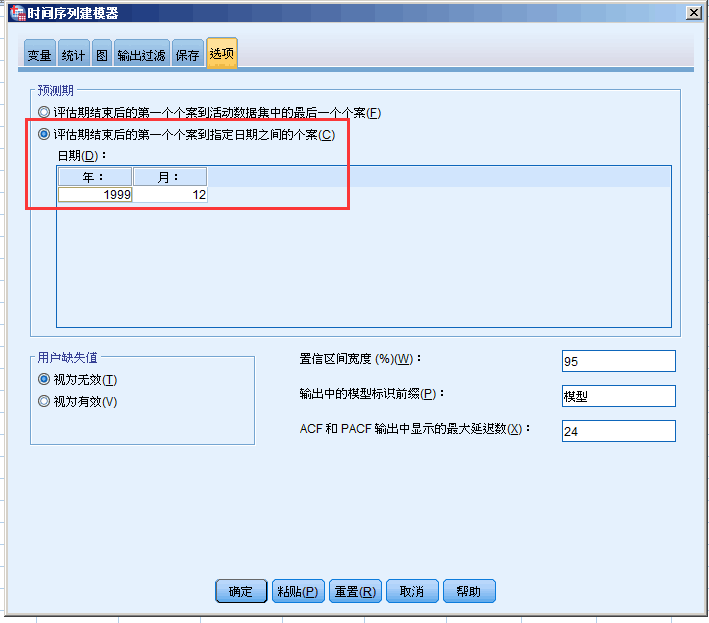
通过分析--时间序列预测--创建传统模型











可以直接在数据集中生成指定时间周期的预测值，预测时间不宜过大

1. 应用时间序列模型

### 【数据分析的思路】

分类变量对分类变量的影响：卡方检验

分类变量对连续变量的影响：t检验/方差分析

连续变量对连续变量的影响：相关分析

得到结论：变量和变量之间有没有显著的影响

为了做影响因素的分析/预测--回归

第一步：分析各个变量自身（分类/连续）

第二步：分析自变量对因变量的影响（分类对分类，分类对连续，连续对连续，连续对分类）

第三步：利用有显著影响的自变量和因变量进行建模（线性回归/Logistic回归）

## 自主学习作业讲解

无

## 课程总结

* 方差分析
* 双因素方差分析
* 卡方检验
* Rxc卡方检验
* 时间序列分析
* 数据分析的思路

## 下次自主学习任务布置

#### 观看预习视频

数据建模分析师——SPSS数据分析

#### 课后作业

1. 参见直播平台
2. Qq群课后习题