

# SPSS统计软件及其应用

Tang Junmin

杭州电子科技大学  
理学院

2011.02



# 提纲



# SPSS课程安排及考核办法

- 一共有16次课，每次两个课时
- 安排理论课7次，上机实习9次
- 前面14次课上课和上机交替进行，作业有7次，少于7次者或完成情况欠佳者酌情扣分（**作业和平时出勤占总成绩的60%**）
- 最后两次安排上机
- 倒数第2次安排复习，讲解考核需要掌握的主要内容
- 最后一次安排上机考试(**上机考试成绩占总成绩的40%**)



# SPSS的历史

- SPSS(Statistical Package for the Social Science) 社会科学统计软件包是世界著名的统计分析软件之一。
- 20世纪60年代，美国Stanford大学的三位研究生研制开发了最早的 统计分析软件，在中小型计算机上运行，后经多年的发展， 推出了可以在电脑微机上运行的pc版，现在的版本是16.0， 并更名为Statistical Product and Service Solutions，即统计 产品与服务解决方案。
- 目前SPSS在全球约有26万家产品用户，分布于通信、医疗、 银行、证券、保险、制造、商业、市场研究、科研教育等多个行业和领域，已经成为世界上最为流行的应用最广泛的专业统计分析软件。



# 统计软件比较之SAS

## STATISTICAL ANALYSIS SYSTEM(SAS)

SAS为美国公司推出的统计分析系统，当前版本为9.1，由于美国政府的支持，SAS软件在美国风行一时，曾有人言，得到SAS软件的认证，你的年薪将不少于20万美金，可是当今世上真正精通SAS的人并不多。

### SAS软件的特点

- 系统巨大，算法繁复，很难知道内部如何运行，以及所得结果的统计含义到底如何？
- 维护费用是一笔不小的支出(SAS公司每年都要收取一定的维护费用，如果不维护，到期无法使用，不单纯是版本低的问题)
- 



# 统计软件比较之SPPLUS/R

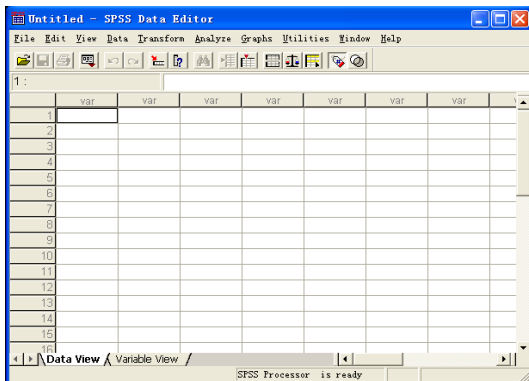
## S+/R

- S+软件是美国bell实验室推出的，当前版本8.0，软件 是建构在S语言之上，可以方便的嵌入在Excel和SPSS中，实现其统计功能，具有灵活，方便，更新速度快，统计过程比较透明的特点，但是 只适合于专业统计工作者使用，多见于国外高校研究机构的科研工作者。
- R软件可以称作是S+的方言版，由澳大利亚的三位统计学家编写，其核心的程序只有30m，其余统计功能包(Package)可以自行下载安装， 源代码全部公开，包的编写有多数专业的统计学家完成，一般包含最 新的统计方法和结果，故可以成为最小，最专业，速度最快的统计软件， 其用户在2006年一年中的增长就超过30万。
- 缺点是**不适合不愿编写程序的人使用。。。。。**



# SPSS的主程序窗口-数据编辑窗口

- 数据编辑窗口



- SPSS的数据编辑窗口(Data Editor)是SPSS的主程序窗口，关闭该窗口就意味着退出SPSS新版16.0可以同时打开多个数据窗口



# 数据编辑窗口功能简介

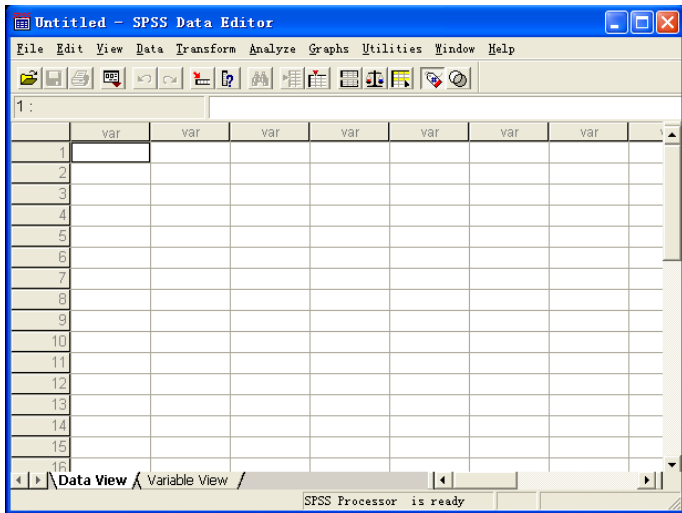
## 数据编辑窗口

- 定义SPSS数据的结构、录入、编辑和管理待分析的数据
- SPSS的所有统计分析功能都是针对该窗口中的数据。
- 数据通常以SPSS数据文件的形式保存在计算机磁盘上，其文件扩展名为“\_.sav”。sav文件格式是SPSS独有的，一般无法通过其他软件如Word, Excel等打开。





# 数据编辑窗口的构成



# 数据编辑窗口的构成

- 窗口主菜单， 菜单的主要功能有文件操作(File)、数据编辑(Edit)、 浏览(View)、数据操作(Data)、数据转换(Transform)、统计分析(Analysis)、制作图形(Graphs)、实用程序(Utillities)、窗口管理(Windows)、帮助(Help)
- 工具栏， 同其他的常用软件一样， SPSS也将一些常用的功能以图形 按钮的形式组织在工具栏中，使用更加便捷和方便
- 数据编辑区， 类似与Exel的表格处理，但是SPSS有两张电子表格，一张处理SPSS数据的结构，另一张处理SPSS的数据
- 系统状态显示区， 系统状态显示区用来显示系统的当前运行状态，当系统等待 用户操作时，会出现"SPSS processor is ready"的提示信息。



# SPSS软件的退出

- 软件退出类似其他窗口操作软件，点击右上角关闭即可
- 但退出软件时会出现几个提示框，提示保存数据或分析结果，或程序语句
- 此时需根据实际情况进行操作，保存需要的程序，数据，结果等



# SPSS软件的三种基本运行方式

- ① 完全窗口运行方式，是指在使用SPSS的过程中，所有的分析操作都通过点击菜单，工具栏按钮，输入对话框等方式完成。它是一种最常见和普遍的使用方式，其最大优点是简洁和直观。这种方式很适合于一般的统计分析人员和SPSS的初学者。
- ② 程序运行方式，是指在使用SPSS的过程中，统计分析人员根据分析的需要，将数据分析的步骤手工编写成SPSS命令程序，然后提交计算机一次运行。这种方式很适用于大规模的统计分析工作(SPSS的高级用户)。
- ③ 混合运行方式，是指在使用菜单的同时还可以编写SPSS程序，是完全窗口菜单方式和程序运行方式的综合。

本课程以菜单操作为主，适当地以程序为辅助。下面简单介绍程序运行方式



# 程序运行方式

程序运行方式很适用于大规模的统计分析工作，此时系统可以依照程序自动进行多步骤地复杂数据分析，分析过程无需人工干预

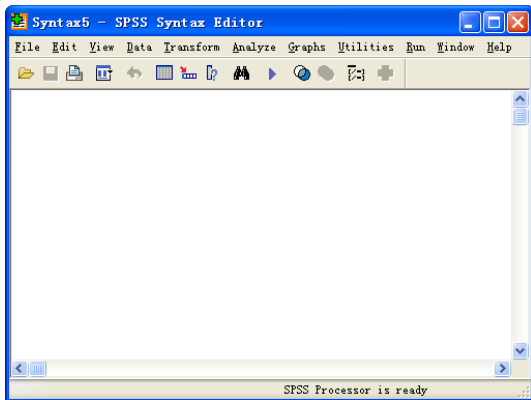
## 程序运行的两项工作

- 1 编写SPSS程序，在Syntax窗口中编辑
- 2 提交并运行SPSS程序



# SYNTAX程序编写窗口

手动点击菜单file → new → Syntax 可得下述窗口， 可以编写程序， 新版15.0中有output(结果)窗口， 在输出结果 之前， 输出的是菜单操作对应的程序， 便于学习程序编写



# SPSS程序执行的四种方式

## 程序执行方式

- ① 全部执行方式(all),依次执行当前语句窗口中的所有SPSS命令程序
- ② 选中执行方式(selection),仅执行当前语句窗口中被选中的SPSS命令程序
- ③ 当前执行方式(Current),仅执行当前语句窗口中当前光标所在行的SPSS命令
- ④ 至末尾执行方式(To end),执行当前窗口, 光标所在行以后的所有的程序命令



# 程序运行方式的注意事项

- 上述四种命令执行方式可以方便程序员调试和运行程序
- 但编写SPSS程序必须按照一定的语法规则，掌握SPSS语法规则，熟悉SPSS命令是编写程序的前提，因此，程序运行方式**一般适合于SPSS的高级用户**，或在SPSS程序员帮助下使用
- **特别提醒，如果需要想学习深入的统计，请学习其他的专业统计软件，不要再SPSS的Syntax语句上花功夫**





# 明确数据分析目标

## 明确数据分析目标

明确数据分析目标是数据分析的出发点，明确数据分析目标就是要明确本次数据分析要研究的主要问题和预期的分析目标等。只有明确了分析目标，才能够正确地制定数据采集方案，如应收集哪些数据，应采用怎样的方式收集等，进而为数据分析做好准备



# 正确收集数据

## 正确收集数据

正确收集数据是指应从分析目标出发，排除干扰因素，正确收集服务于既定分析目标的数据，正确的数据对于实现数据分析目的将起到关键性的作用。

数据收集设计到抽样调查等理论，简单一点讲，需要采用随机分组，排除干扰因素，尽可能得到净化的数据。因此，排除数据中那些与目标不关联的干扰因素是数据收集中的重要环节。



# 数据的加工整理

## 数据的加工整理

在明确数据分析目标的基础上收集到正确的数据，需要进行必要的加工处理后才能真正用于统计分析建模。通过数据的加工整理，人们能够大致掌握数据的总体特征，是进入深入分析的基础。数据加工整理通常包括数据的缺失值处理，数据的分组，基本描述统计量的计算，基本统计图形的绘制，数据取值的转换，数据的正态化处理等。



# 明确方法，正确分析

## 明确统计方法的含义和适用范围

数据加工整理之后，一般可以进行统计分析了。统计分析时切忌滥用和误用统计分析方法。滥用和误用的主要原因是对于方法解决哪类问题，方法适用的范围，方法对数据的要求不清楚。另外，统计软件的不断普及和应用中的不求甚解也会加重这些现象。因此在数据分析中应该避免盲目的“拿来主义”，否则分析的结果可能出现较大的错误



# 读懂结果，合理解释

## 读懂分析结果，正确解释分析结果

数据分析的直接结果是统计指标和统计参数。正确理解这些指标和参数的统计含义是一切分析结论的基础。正确理解统计指标和统计参数的含义是证实分析结论正确性和可信性的依据，这一切都取决于人们 能否正确地把握统计分析方法的核心思想。



# 利用SPSS进行数据分析的一般步骤

- ① 数据准备阶段：根据SPSS的要求，利用SPSS提供的功能准备SPSS数据文件，其中包括定义数据的结构，录入修改数据
- ② 数据加工整理阶段：对SPSS数据编辑器窗口中的数据进行必要的预处理
- ③ SPSS数据分析阶段：选择正确的统计分析方法，对数据编辑器窗口中的数据进行统计分析
- ④ SPSS分析结果的阅读和解释：读懂SPSS输出窗口中的分析结果，明确统计含义，并结合应用背景知识做出切合实际的解释

以后我们学习过程中也遵循这样的步骤，先分析数据，再预处理，进行统计分析，解释统计结果



# SPSS数据文件的特点

## SPSS数据文件的特点

- SPSS 数据文件的扩展名是.sav
- SPSS数据文件是一种有结构的数据文件

基于上述特点，建立SPSS数据文件时应完成两项任务

- 描述SPSS数据的结构
- 录入编辑SPSS数据



# SPSS数据的基本组织方式之一

SPSS的数据将直观地显示在数据编辑窗口中，形成一张平面二维表格。待分析的数据有两种方式组织在表格中。

## 原始数据的组织方式

如果待分析的数据是一些原始的调查问卷数据，或是一些基本的统计指标，那么这些数据就应以原始数据的组织方式组织。在这种组织方式中：

- 数据编辑窗口的每一行称为一个个案，或称为一次观测，所有个案组成了SPSS数据文件的内容。
- 数据编辑窗口的每一列称为一个变量，变量都有一个名字，称为变量名，它是访问和分析SPSS每个变量的惟一标志。
- SPSS数据文件的结构就是对每个变量及相关特征的描述





## 数据基本组织方式之原始数据举例

zgh	xb	nl	sr	zc	xl	bx
001	男职工	48	1014.00	高级工程师	本科	12.00
002	男职工	49	984.00	工程师	专科	9.00
003	男职工	54	1044.00	高级工程师	高中	13.00
004	男职工	41	866.00	助理工程	高中	8.00
005	男职工	38	848.00	助理工程	本科	8.00
006	女职工	41	824.00	无技术职	高中	7.00
007	女职工	42	824.00	无技术职	高中	7.00
008	女职工	41	824.00	无技术职	高中	7.00



# SPSS数据基本组织方式之二

## 频数数据的组织方式

如果待分析的数据不是原始的调查问卷数据，而是经过分组汇总后的 汇总数据，那么这些数据就应该以频数数据的组织方式组织。

- 数据编辑窗口中的一行为变量的一个分组（或多个变量交叉分组下的一个分组）
- 所有行囊括了该变量的所有分组情况（或多个变量交叉下的所有分组情况）
- 数据编辑窗口的一列仍为一个变量，代表某个问题或方面以及频数



## 频数数据的组织方式示例

		1	2	3
		zhicheng	nianling	renshu
1		1	1	0
2		1	2	15
3		1	3	8
4		2	1	10
5		2	2	20
6		2	3	2
7		3	1	20
8		3	2	10
9		3	3	1
10		4	1	35
11		4	2	2
12		4	3	0



# SPSS变量名

变量名是变量访问和分析的惟一标志.在定义SPSS数据结构的时候应首先给出每列变量的变量名

## 变量的命名规则

- 变量名总长度不能超过8个字符
- 首字符应以英文字母或汉字开头，后面可以跟除了!、?、\*之外的字母或数字，不能以下划线、圆点为终止符号。
- 变量名不区分大小写字母。允许汉字作为变量名，汉字总数不能超过四个
- 变量名不能与SPSS内部特有的具有特定含义的保留字符同名，如ALL、BY、AND、NOT、OR等
- 总之，在为变量命名时，为方便记忆，变量名最好与其代表的数据含义相对应。
- 也可以使用变量名标签来更好地解释变量的数据含义



# 数据类型(TYPE)

数据类型是指每个变量取值的类型，SPSS中有三种基本数据类型，分别为数值型，字符型和日期型。

- 数值型(numeric): 用于描述纯粹数值型变量和可数值化的变量，如性别，年龄，职称，学历，工资，成绩等
- 字符型(string): 描述字符型变量，如职工号码，姓名，地点，建议，备注等无法分类的字符串
- 日期型(data): 用来表示日期或者时间，如生日，成立日期等变量。有很多格式可选



## 变量名标签(VARIABLE LABEL)

变量名标签是对变量名含义的进一步解释说明，它可增强变量名的可视性和统计分析结果的可读性。变量名标签可用中文，总长度可达120个字符，但在统计分析结果的显示中，一般不可能显示如此长的变量名 标签信息。变量名标签属性可以省略，但最好是给出变量名的标签。



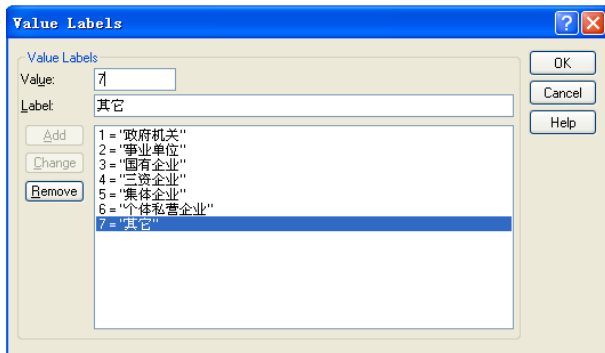
## 变量值标签(VALUE LABEL)

变量值标签是对变量取值含义的解释说明信息，对于品质数据或类别数据尤其重要，例如对于性别变量，可以用1表示男，2表示女，但是用户看到的只是1和2这样的数据，如果给性别变量附加变量值标签，并给出1和2的实际指代，则可以更清楚地表达数据的含义。

通常，变量值标签对于顺序水准(如收入的高、中、低)和名义水准(如民族、性别)的品质型变量来说是必不可少的。它不仅明确数据含义，也增强了最后分析结果的可读性



# SPSS中定义变量值标签的窗口



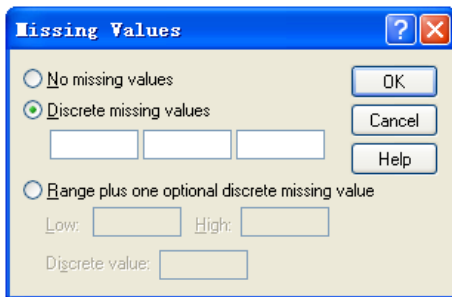


## 缺失数据(MISSING)

缺失数据的处理时数据分析准备过程总的一个重要的环节，在统计分析过程中收集来的数据可能出现以下问题

- 数据中存在明显错误或明显不合理的数据
- 数据中存在漏填数据项

统计学中常称上述情况的数据为不完全数据或缺失数据。SPSS中说明缺失数据的基本方法是指定用户缺失值。定义窗口如下



# 变量的度量尺度

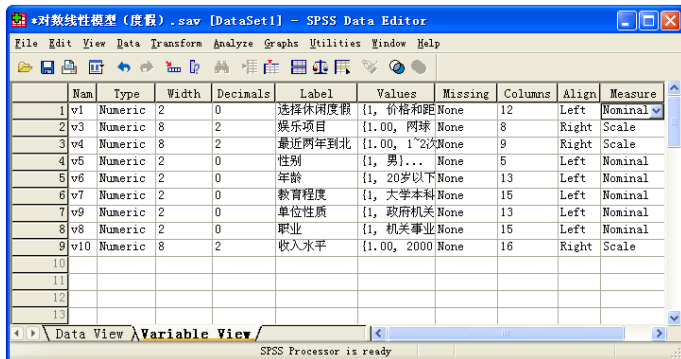
## 度量尺度的三大类

- 定距型数据(Scale): 通常是指诸如身高, 体重, 年龄, 血压, 存(取)款的金额等连续型数据, 也包括诸如人数, 商品件数等离散型数据。它不能是字符型(string)变量, 但可以是数值型(numeric)或日期型(date)。
- 定序型数据(Ordinal): 该类数据具有内在固有大小或高低顺序, 如职称, 年龄的分段, 收入水平的分段, 质量等级等。它可以是字符型(string)变量, 也可以是数值型(numeric)。
- 定类型数据(Nominal): 该类数据没有内在固有的大小或高低顺序, 一般以数值或字符表示的分类数据。例如, 性别变量中的男女取值, 可以分别用1和2表示。它可以是数值型(numeric), 也可以是字符型(string)变量, 例如姓名。



# 结构定义的基本操作

定义SPSS数据结构的操作是在数据编辑窗口进行的。数据编辑窗口有两个卡片，为Data View和Variable View，其中Variable View就是专门用来完成数据结构定义的，如下图



我们可以根据上面变量的各种属性定义自己需要的变量



# 变量定义举例

## EXAMPLE (调查问卷)

(请在选项上打勾)

你的性别 : (1) 男 (2) 女

年级 : (1) 2004级 (2) 2005级

(3) 2006级 (4) 2007级

一、你认为目前学习中存在的最大问题是

A 不喜欢所学专业, 学习兴趣不浓

B 学习方法不科学, 效率不高

C 所学内容过于枯燥、陈旧, 负担太重

D 只掌握了书本知识, 缺乏实践能力

我们应该如何定义变量来表述上述调查问卷中的问题?



# 设计的变量

## EXAMPLE (性别)

变量名 : 性别  
数据类型 : `numeric`  
变量值标签: 1 男  
              : 2 女  
`measurement: nominal`  
.....

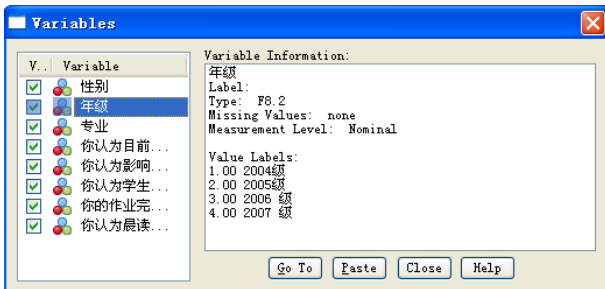
## EXAMPLE (年级)

变量名 : 年级  
数据类型 : `numeric`  
变量值标签: 4 2004  
              : 5 2005  
              .....  
`measurement: ordinal`  
.....



# 变量属性列表浏览

SPSS数据的结构定义完成之后，一般会希望浏览一下整个数据文件中所有变量的特征和属性，可以通过菜单选项 **Utilities→variables** 来实现，弹出如下窗口



# SPSS数据的录入

SPSS数据的录入操作时在数据编辑窗口-Data View中进行的，操作方法与Excel相似，注意以下几点

- 黑框[ ]所在的单元为当前数据单元，可以录入和修改
- 用tab键切换到右边，一行结束时tab键可切换到下一条数据,按Enter键切换到下方单元格;
- 变量值标签数据可以通过下拉按钮完成，但之前应该打开变量值标签的显示开关， 点击view→Value labels。



# SPSS数据的编辑

## 数据编辑的主要操作

- 数据的定位
  - 按个案号码自动定位 ,Edit→Go to case
  - 按变量值定位, 选中变量, Edit→Find
- 插入和删除一个个案, 一行, 选中行, 右键, cut
- 插入和删除一个变量, 一列, 操作同上
- 数据的移动、复制和删除

- 具体操作看演示






# SPSS数据的存储

## SPSS支持的数据格式

- SPSS 格式文件，以.sav作为扩展名，可以为SPSS直接读取，但通用性较差
- Excel 文格式件，是应用十分广泛的电子表格文件，扩展名为.xls
- dbf 格式文件，使一种应用较为广泛的数据库格式文件，其扩展名为.dbf
- 文本格式文件，扩展名为.dat，第一行为变量名，从第二行起是数据
- 上述后三种数据文件的优点是可以方便的被软件读取，缺点是只保存相应的变量值，而没有保存变量属性，有利于统计分析的必要信息会丢失



# 保存SPSS数据的基本操作

- 选择菜单File→Save as，出现 
- 给出存放数据文件的目录路径和数据文件的文件名
- 数据文件的格式在"保存类型"的下拉框中选择
- 点击variables按钮，选择需要保存的变量
- 点击保存即可



# SPSS可以直接读取的数据文件格式

- SPSS格式文件,扩展名.sav
- Excel 格式文件, 扩展名.xls
- dBase系列数据文件,扩展名为.dbf
- SAS格式文件, 扩展名为.sas7bdat

## 基本操作步骤为

- 选择菜单File→Open→Data
- 选择数据文件的类型,并输入数据文件名



# SPSS打开其他数据格式文件的注意事项

- 如果读入的是dBase数据文件，数据库文件的的字段名，字段类型将自动转成SPSS中的变量名和变量类型， 其中的一个记录转成SPSS的一个个案
- 如果读入的是Excel格式文件， SPSS默认将Excel工作表中的全部数据读到SPSS的Data View编辑窗口中，也 可以读取指定区域中的数据，工作表上的一行读取为一个个案，如果Excel工作表的第一行是变量，则应选择Read variable names项，进行特别处理



- 文本数据文件的读取见



# 数据的预加工处理

## 数据的预加工处理

数据文件建立之后，通常需要对所分析的数据进行必要的预加工处理，这是数据分析过程中不可缺少的步骤，而且随着数据分析的不断深入，对数据还要进行多次反复加工处理。

- 缺失值和异常数据的处理。
- 数据的转换处理。在原有数据的基础上，产生一些含有更丰富信息的新数据，或对数据的原有分布进行转换等
- 数据抽样。按照一定的规则从大量的数据中选取部分样本进行分析
- 选取变量。不是所有变量都是有意义的，故而选取部分变量进行分析是很自然的



# 数据的排序

数据排序是将数据编辑窗口中的数据按照某个或多个指定变量的变量值升序或降序进行重新排列,此处的变量称为排序变量。

排序在数据分析中的作用有如下几点

- 数据排序便于数据的浏览
- 通过排序很容易找到数据的范围,可以初步把握数据的离散程度
- 通过排序可以便捷地找到数据的异常值,便于进一步的分析

说明 (1)数据排序是整行数据排序,而不是只对某列变量进行排序;(2)数据排序后,原有数据的排序次序必将打乱,因此在数据排序前,应注意保留数据的原始排序顺序,以免发生混乱。



# SPSS数据排序的基本操作

- 选择菜单Data→Sort cases,弹出下述窗口
- 将主排序变量从左边的列表框中选到sort by框中，并在sort order中选择升序还是降序
- 如果是多重排序，还需要依次指定第二、第三排序变量及相应的排序规则。



# 数据排序的应用举例

## 对居民储蓄数据进行如下排序

- 分别对城市和农村关于一次存款金额进行升序排序
- 分别对男性女性关于年龄进行降序排序





# 数据文件的合并

SPSS提供两种合并数据文件的方式，分别是纵向合并和横向合并。

## 纵向合并数据文件

纵向合并数据文件就是将数据编辑窗口中的数据与另一个数据文件进行首尾对接，将另一个数据文件的内容追加到当前数据文件之后，并依据两个数据文件中的变量名进行数据拼接。效果是增加个案，所以又称之为个案合并。

## 横向合并数据文件

横向合并数据文件就是将数据编辑窗口中的数据与另一个SPSS数据文件中的数据进行左右拼接，即将一个SPSS数据文件的内容拼接到数据编辑窗口中当前数据的右边，并依据两个文件的个案进行数据对接。效果是增加观测变量，所以又称之为变量合并。

数据拼接以职工数据为例，具体操作见



# 变量计算的目的

变量计算是数据分析过程中应用最为广泛的一个关节，通过变量计算可以处理许多问题

- 数据的转换处理:指在原有的数据基础上，计算产生一些含有更丰富信息的新数据或感兴趣的数据指标。
- 对数据的原有分布进行转换:由于数据分析和建模的需要，对数据的分布有一定的要求，因此可以利用变量计算对原有数据的分布进行变换，例如利用对数或多项式对非正态或非线性数据进行处理，对事件序列进行平稳化处理等



# SPSS变量计算

SPSS变量计算是在原有数据的基础上，根据用户给出的SPSS算术表达式以及函数，对所有个案或满足条件的部分个案，计算产生一系列新变量。

- 变量计算是针对所有个案的，每个个案都有自己的计算结果
- 变量计算的结果应保存在一个指定的变量中，该变量的数据类型与计算结果的数据类型相一致

在变量的计算过程中涉及到SPSS的几个概念：SPSS算术表达式，SPSS条件表达式，SPSS函数



# SPSS算术表达式

NUMERIC EXPRESSION是由常量、变量、算术运算符、圆括号、函数等组成的式子

- 字符型常量应当用引号""括起来
- 变量是指那些已经存在于数据编辑窗口中的原有变量
- 算术运算符主要包括+、-、\*、/、\*\*(乘方)
- 在同一算术表达式中的常数及变量，数据类型应一致，不然无法计算。



## SPSS条件表达式

条件表达式是一个对条件进行判断的式子，包括简单和复杂两种条件表达式，其返回的结果是真或假

- ① 简单条件表达式由关系运算符、常量、变量以及算术表达式构成，其中常用的关系运算符有  $>$ ,  $<$ ,  $=$ ,  $\neq$  (不等于),  $\geq$ ,  $\leq$
- ② 复合条件表达式也称为逻辑表达式，是由逻辑运算符、圆括号和简单条件表达式等组成。其中常用的逻辑运算符包括  $\&$  或 AND (逻辑且)、 $|$  或 OR (逻辑或)、 $\sim$  或 NOT (逻辑非)，NOT 的运算级别最高，其次是 AND，最低是 OR，必要的时候可以通过圆括号改变运算的优先级。



# SPSS函数

SPSS函数是事先编好并存储在软件中，能够实现某些特定计算任务的计算机程序，函数都有自己特定的函数名，函数名的第一个字符一般为大写，根据函数功能和处理的变量类型，SPSS函数大致分为八大类，见下页



# SPSS函数的分类

**算术函数** 主要完成一些特定的算术计算功能，函数值和参数通常为数值型

**统计函数** 主要计算基本描述统计量，函数值和参数为数值型

**分布函数** 用来产生一个服从某种统计分布的随机数，函数值为数值型

**逻辑函数** 进行逻辑判断

**字符函数** 对字符型数据进行处理

**日期函数** 对日期进行处理

**缺失值函数** 用于判断缺失值

**其他函数** 除上述函数外，SPSS还有一些辅助函数。



# 变量计算的基本操作

- 打开菜单Transform，选择Compute variable
- 在target框中输入存放计算结果的变量名，可以使用新的变量，也可以覆盖原有变量，变量属性可以在Type&Label中修改
- 在Numeric Expression中给出SPSS算术表达式和函数，可以手工输入，也可以点击完成
- 如果用户希望对符合某种条件的个案进行变量计算，可以单击If,输入条件表达式，
- 具体操作





# 数据选取的目的

数据选取就是根据分析的需要，从已收集的数据 中按照一定的规则抽取部分数据参与分析的过程，通常也称为 样本抽样。其主要目的有

- ① 提高数据分析效率，如果数据量较大，会在一定程度上影响计算 和建模的效率，因此抽出少量样本进行操作，可以提高分析的效率
- ② 检验模型的需要，在数据分析中，所建模型是否能够较完整准确地反映数据的特征，是否能够用于以后的数据预测，这些问题都是人们极为关心的，为了验证模型，一般可依据一定的抽样方法只选择部分 样本参与数据建模，而剩余的数据用于模型检验



## 数据选取的基本方式

- ① 按指定条件选取，即选择符合条件的数据，这里SPSS要求用户以条件表达式的形式给出数据选取的条件。
- ② 随机抽样，对数据编辑窗口中的所有个案进行随机筛选，有如下两种方式
  - ① 近似抽样，要求用户给出一个百分比数值，SPSS按照这个比例自动从数据编辑窗口中随机抽取相应百分比的数据，但抽取的数据不一定恰好等于要求的百分比，有一定的偏差
  - ② 精确抽样，要求用户给出两个参数，第一个参数是希望选取的个案数，第二个参数是指定在前几个个案中选取



# 数据选取的基本操作

## 操作步骤

- ① 选择菜单 Data→Select cases, 弹出窗口
- ② 根据需要选择数据的选取方法
- ③ 指定对未选中个案的处理方式, 其中Filtered表示在未选中的个案号码上打一个"/"标记, Deleted表示将未选中的 个案从当前窗口删除。
- ④ 具体操作
  - 指定条件选取数据, 只分析城镇储户的情况
  - 近似抽样, 抽出70%的数据进行分析



# 计数的目的和要求

计数对于把握个案个方面的特征十分有效。SPSS实现的计数是对所有个案或满足条件的部分个案，计算若干变量中有几个变量的值落在指定的区间中，并将计数的结果存入一个新变量中的过程。如分析大学毕业班中同学成绩时，可以依次分析每个同学若干门专业课中优、良、中、及格、不及格的门数。

SPSS实现技术的关键步骤是：

- ① 指定哪些变量参与计数，计数的结果存入哪个新变量中
- ② 指定计数区间，尤其关键



# 计数区间



计数区间是个广义的概念，可以有以下描述形式

- ① 单个变量值(Value)
- ② 系统缺失值(System-missing)
- ③ 系统缺失值或用户缺失值(System or user-missing)
- ④ 给定最大值和最小值的区间(n through m)
- ⑤ 小于等于某给定值的区间(Lowest through n)
- ⑥ 大于等于某给定值的区间(n through highest)

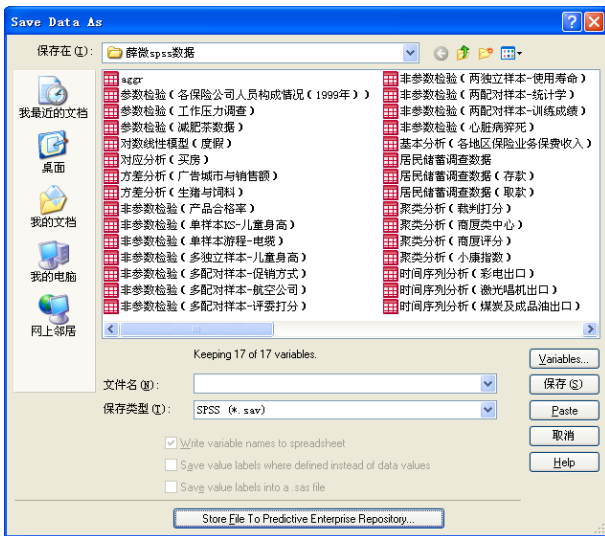


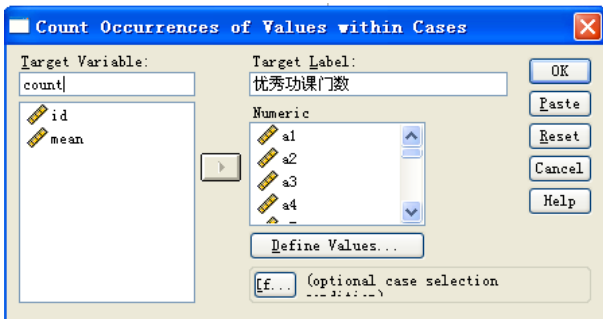
# 计数的基本操作

## 操作步骤

- ① 选择菜单 Transform → Count values within cases, 弹出 
- ② 将参与计数的变量选到Numeric variables框中。
- ③ 在Target Variable框中输入存放计数结果的变量名，并在Target Label中输入变量标签
- ④ 点击 Define Values按钮定义计数区间，出现 
- ⑤ 点击add, chang, remove等按钮完成计数区间的增加，修改和删除
- ⑥ 若只对满足条件的个案进行计数，可以单击If按钮输入相应的条件表达式







◀ return





Count Values within Cases: Values to Count

Value

☐ Value:

☐ System-missing

☐ System- or user-missing

☒ Range:

through:

☐ Range, LOWEST through value:

☐ Range, value through HIGHEST:

Add

Change

Remove

Values to Count:

90 thru Highest

Continue Cancel Help

◀ return



# 分类汇总

分类汇总是按照某分类方法进行分类汇总计算，在实际数据分析中非常常见


分类汇总主要涉及两个方面

- 按照哪个变量进行分类
- 对哪个变量进行汇总，并指定哪些统计量



# 分类汇总的基本操作

## 操作步骤

- 选择菜单Data→Aggregate，出现窗口 
- 将分类变量选到 Break Variable框中
- 将汇总变量选到 Aggregate Variable中
- 单击 Function按钮，指定对汇总变量计算哪些统计量（默认为计算均值）
- 指定将分裂汇总结果保存到何处。可选
  - Create new data file 表示保存为新的数据文件
  - Replace working data file 表示覆盖当前数据编辑窗口中的数据
- 单击name&Label 按钮，重新指定结果文件中的变量名或添加变量名标签。



# 分类汇总的应用举例

## EXAMPLE

试对居民储蓄调查数据，进行如下分析 分析城镇储户和农村储户的一次平均存(取)款金额是否有显著的差异。这里可进行的初步分析是按照户口类型对存(取) 金额进行分类汇总，其中分类变量是户口，汇总变量是存(取)金额，计算其均值和标准差。



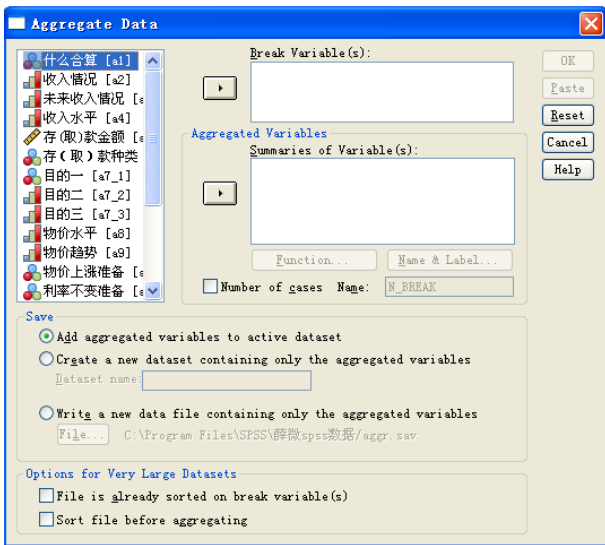


FIG: 数据分类汇总

## 你想过下面的问题吗？

- ① 当你买一台洗衣机时，被告知三年免费保修。厂家凭什么这么说呢？如果说的保修时间多了，那么厂家就会亏；如果说得少了，产品就没人买，也是损失。那么这个保修期到底是怎样确定的呢？
- ② 大学排名是一个非常敏感的话题，不同的机构会得出不同的结果，各自都说自己的结果是客观、公正和有道理的。到底如何理解这些不同的结果呢？
- ③ 任何公司都有个信用的问题。如果这些公司试图得到贷款时都没有不还贷的不良记录，那么如何根据他们的财务和商业资料来判断其信用等级呢？
- ④ 疾病传播时，如何能够通过被感染者入院前后的各种经历得到一个疾病传染方式的模型呢？



- ① 如何通过问卷调查来得到性别、年龄、职业、收入等各种因素与公众对某项事务(比如商品或政策)的态度的关系呢?
- ② 一个从来都没有研究过红楼梦的统计学家如何根据写作习惯得出红楼梦从哪一段开始就不是曹雪芹的手笔了呢?
- ③ 怎样才能够客观地得到某个电视节目的收视率, 以确定插播的广告价格是否合理呢?

其实, 这些都是统计应用的例子。这样的例子太多了, 因为统计学可以应用到几乎所有的领域, 包括保险精算, 农业, 动植物学, 经济计量学, 流行病学, 金融, 遗传学, 历史研究, 工业, 法律, 文学, 管理科学, 市场营销学, 气象学, 心里学, 质量控制, 社会学, 抽样调查, 博彩等。当然, 我们不可能理解所有的统计应用, 只要能够解决身边的统计问题就够了。



# 现实生活中的随机性和规律性

我们学过了很多的科学定律，例如牛顿三大定律，物质不灭定律，以及化学中的很多定律。但是在很多领域，很难用如此确定的公式 或论述来描述一些现象。

## EXAMPLE (人的寿命)

人的寿命是很难预先确定的。一个经常吸烟、喝酒、不锻炼身体而且喜好油腻食物的人可能比一个很少得病、生活习惯良好的人活得长。因此，可以说，人的寿命的长短具有一定的随机性。这种随机性可能和人的经历、基因、生活习惯等许多不易说清的因素有关。





# 寿命的规律性

## EXAMPLE (寿命中的规律性)

从总体上说，我国公民的预期寿命是非常稳定的，而且由于生活水平的提高，寿命也在逐步增长，如1996年的平均预期寿命是70.8岁，而2000年时为71.4岁。这就是规律性。一个人可能活过这个平均年龄，也可能活不到这个年龄，这是随机的。但总体上说，预期寿命的稳定性，说明了随机中的规律，这个规律就是统计规律。



# 变量

- 做任何事情都有对象,比如上SPSS课的同学,一共是60人,这是固定的,称为**常数**。
- 但是要考虑今天上课的人数,那就不确定了,具有随机性,可能有请假的、无故逃课的,等等。这时上课的人数就是个变量(variable)。
- 另外,对于某项政策同意与否的回答,有“同意”、“不同意”、“不知道”三种可能值,这也是变量,只不过不是数值而已。
- 变量的划分
  - 如果变量按随机规律所取的值是数量值,则称为定量变量
  - 如果变量的取值非数量值,则称为定性变量或属性变量或分类变量



# 数据

有了变量的概念，那么什么是数据呢？拿掷骰子来说吧，掷一次骰子会得到什么值，这是一个随机变量。每次取得1至6点中任意某点数的概率在理论上都是六分之一，而在实际掷骰子的过程中，如果掷100次，会得到100个从1到6点组成的数字串，再掷100次，又得到一个数字串，和前一次的结果多半不一样。这些试验的结果就是数据。所以说：**数据是变量的观测值**



# 这个世界是充满未知关系的世界，我们需要探索的东西很多

现实世界中的各种问题都是相互联系的。不讨论变量之间的关系，就无从谈起任何有深度的应用；而没有应用，统计的基本概念就仅仅是个摆设。

人们时刻都在关心事物之间的关系，比如政府十七大与现在的股市大盘、毕业后的职业和收入的关系、广告投入和经济效益之间的关系、治疗手段和治愈率之间的关系等等。这些都是二元关系，还有其他更加复杂的诸多变量之间的相互关系，如企业的固定资产、流动资产、预算分配、管理模式、生产率、债务和利润等诸因素的关系是不能用简单的一些二元关系所描述的。



# 定量变量间的关系

考虑广告投入和销售之间的关系

TABLE: 某企业广告投入和销售 额原始数据(万元)

广告	1.0	3.2	3.3	5.5	5.9	7.1	7.3	9.2	10.8	12.1
销售	9.4	31.8	33.2	52.4	53.5	56.0	56.9	59.2	60.1	63.5

广告投入和销售 额之间到底有没有关系呢?我们可以用二维散点图来感受一下这个数据，

► 散点图



# 广告投入和销售额之间的关系

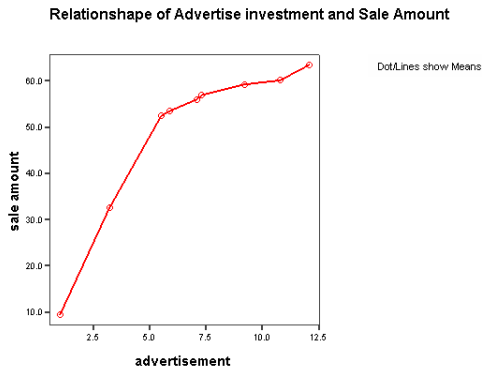


FIG: 广告投入和销售额之间的关系

## 一般地，人们希望能够从数据回答几个问题

- ① 这两个变量之间是否有关系？显然，他们有关系，从散点图很容易看出，销售额随着广告投入的增加而增加
- ② 如果有关系，那么它们的关系是否显著？这个也可以从散点图得到。当广告投入在6万元以下时，销售额增长很快，但是大于这个投入后，销售额增长就不明显了。因此这两个变量的关系是由强到弱。
- ③ 这些关系是什么关系，是否可以用数学模型来描述？本例看上去可以拟合一个回归模型，但绝不是线性的
- ④ 这个关系是否具有普遍性？对其他的企业也是对的吗？现有的数据还不足以回答这个问题，需要考虑更多的变量和收集更多的数据



# 定量关系与因果关系

- 关于因果关系
- 在可控制的试验中，较容易找到因果关系；比如治疗方式和疗效的关系等
- 但是，一般来说，变量之间有关系这个事实并不意味着一定存在明确的因果关系。

## EXAMPLE

- 比如，北京GDP在一年中是快速增长的，而一个刚出生的巴拿马婴儿在这一年中的体重也是快速增长的
- 如果画出图来，它们有类似线性的关系
- 但它们显然没有因果关系





# 定量关系与因果关系

- 只要有关系，即使不是因果关系也不妨碍人们利用这种关系来进行推断。
- 比如利用公鸡打鸣来预报太阳升起；虽然公鸡打鸣绝对不是日出的原因(虽然打鸣发生在先)
- 简单的办法（诸如画图）可以得到一些信息，但不一定能够给出满意的答案
- 需要更多的工具和手段来进行数值分析得到更加严格和精确的解答
- 所以还需要学习我们的统计分析课程



# 定性变量之间的关系

## EXAMPLE

下面是对123人进行关于某项政策调查所得结果的一个简单的三维表，它显示了人们的收入和性别对该项政策的观点。

性别	观点：反对			观点：赞成		
	低收入	中收入	高收入	低收入	中收入	高收入
男	5	8	10	20	10	5
女	2	7	9	25	15	7

FIG: 不同收入和不同性别人群对某项政策的观点



# 定性变量之间的关系

- 从这个数据，希望可以看出收入、性别对观点是否有影响及如何影响
- 如果要得到更加精确的结论，就要进行进一步的分析和计算
- 这是后面交叉列联表分析或多项分布对数线性模型的内容



# 定性和定量变量间的混和关系

- 有些数据不是仅有定性变量或仅有定量变量
- 需要知道包括定性和定量两种变量的一些变量之间的关系
- 下面数据包含两种变量,假象数据的变量是关于高等学校的一些指标,
- 指标包括学校名(U),在校学生人数(S), 研究生比例(G),教师人数(F),职工人数(ST),SCI和SSCI文章数目(P),SCI和SSCI文章引用数目(Q),科研项目数 (PR), 科研经费(B),招生范围 (N)



# 定性和定量变量间的混和关系

TABLE: 有关各高校指标的 数据 形式

U	S	G	F	ST	P	Q	PR	B	N
学校1	15000	0.52	2200	6000	3200	200	1000	1000	A
学校2	2000	0.31	3100	8000	1200	80	1200	800	A
学校3	8000	0.42	2000	6500	2100	120	800	500	B
学校4	12000	0.45	1700	5000	1050	60	200	70	B
...	...	...	...	...	...	...	...	...	...



# 定性和定量变量间的混和关系

除了校名外，该数据有9个变量，其中招生范围为定性变量，其余为定量变量。它的每一列为一个变量的观测值，每一行代表一个学校的各个变量的观测值。从这个数据很难马上看到任何关系。但是经过分析，从中可以得到很多有用的关系和结论，比如得到一个变量和其余变量之间的定量关系，也可以利用其中一些变量把各个高校进行分类，还可以作为学校排名的根据。



# 多选项数据分析

## 什么是多选项问题

在回答某个问题时，答案为两个以上

### EXAMPLE

比如：高考填报志愿（可以在一批志愿中选择多个）

- 北京大学
- 清华大学
- 中国人民大学
- 北京师范大学
- .....

多选项问题不能进行直接处理



# 多选题问题的基本编码方法

**二分法** 当所设计的多选题选择项不多的情况下，通常采用二分法对多选题进行编码。基本方式是：多选题有多少个选择项就设计成多少个SPSS变量，每个变量的取值非 0 即 1。

**分类法** 当所设计的多选题选择项很多时，比如选择项有 10 项以上，如果按照二分法编码就应该设计 10 个以上的变量，这时候显得过于繁琐，所以应该选择分类法。基本方式是：首先应估计多选题问题最多可能出现的答案个数，然后为每个答案设置成一个SPSS变量，变量取值为多选题问题中的可选答案。





## 二分类方法举例

### 分析:

此处有四个选项，所以SPSS应相应设计四个变量，可计为 $x_1$ (方法1)， $x_2$ (方法2)， $x_3$ (方法3)， $x_4$ (方法4)，若选择了其中某项，则该变量的取值为1，否则为0。这样将调查问卷的所有记录输入后，依次可得到四个变量的所有取值。



# 多重分类法举例

## EXAMPLE

你最喜欢的矿泉水品牌是 ( ) 请选择五项

1、农夫山泉 2、怡宝 3、统一 4、脉动 5、娃哈哈  
6、乐百氏 7、大峡谷 8、景田 9、... 10、...

### 分析:

此时处理方式是，由于需要列出五项，所以应设计 5 个变量，分别为 BRAND1、 BRAND 2、 BRAND 3、 BRAND 4、 BRAND 5。每个变量的取值为选择项对应的编码。

比如第一个受访者选择了 1、 2、 5、 6、 8，则 BRAND1 = 1， BRAND 2 = 2、 BRAND 3 = 5， BRAND 4 = 6， BRAND 5 = 8。这个受访者可能只喜欢这五种品牌，对其他品牌无所谓。



## 定义多选题变量集，进行多选项分析

完成上述编码过程后，问卷数据变成了便于分析的统计数据，要完成对多选题问题的分析，必须定义多选题变量集。

### EXAMPLE

确诊高血压后，您按医生的建议采取了哪些非药物控制措施？

- 1、调整饮食
- 2、做适量运动
- 3、保持情绪稳定
- 4、其他措施

### 分析:

这个多选题涉及到四个变量，要完成对这个多选题问题的回答，必须将四个变量综合起来才能进行分析，即需要定义多选题变量集。



# 定义多选项变量集的操作

analyze→multiple response→define sets

弹出如下的对话框:

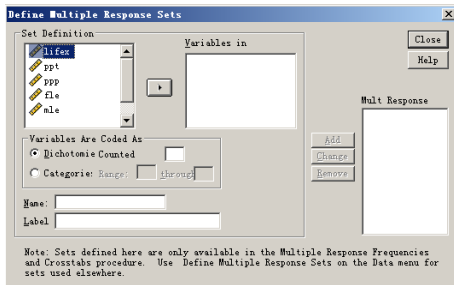


FIG:



# 比率分析

比率分析用于对两变量间变量值比率变化的描述分析，适用于定距型变量（Scale），是SPSS11.0版新增的方法。例如，根据1999年各地区保险业务情况的数据，分析各地区财产保险业务的保费收入占全部业务保费收入的比例情况。通常的分析可以生成各个地区财产保险业务的保费收入占全部业务保费的比率变量，然后对该比率变量计算基本的描述统计量（如均值，中位数，标准差等），进而刻画比率变量的集中趋势和离散程度。比率分析不仅可以实现上述功能，还可以计算其他的相对比描述指标



# RATIO 分析计算的其他的比率指标

- AAD(Average Absolute Deviation)平均绝对偏差

$$AAD = \frac{\sum |R_i - M|}{N}$$

其中  $R_i$  为比率数,  $M$  是比率变量的中位数,  $N$  为样本数。

- COD(Coefficient of Dispersion)离散系数

$$COD = \frac{\frac{\sum |R_i - \bar{R}|}{N}}{M}$$

- COV变异系数

$$COV = \frac{\sqrt{\frac{\sum (R_i - M)^2}{N}}}{M}$$

上述指标从不同的角度刻画了比率变量的集中趋势和离散程度



# 描述性统计分析主要窗口

## 12 描述性统计分析

- 频数分析
- 描述性分析
- 探索性数据分析
- Crosstabs 列联表(交叉表)分析
- Ratio Statistics 比率分析
- ]
- Q-Q图



# 描述性统计分析菜单简介

菜单Analysis 选项 Discriptive statistics(描述性统计)下有6个功能, 分别是

- Frequencies.....(频数分析)
- Discriptives.....(描述性分析)
- Explores....(探索性数据分析)
- Ratios....(比率分析)
- Crosstables....(列联表分析)
- P-P Plots....(Points to Points图)
- Q-Q Plots....(Quantile to quantile 图)

下面分别阐述各个选项的统计分析功能, 其中pp图和qq图暂不作详细介绍





# 频数分析

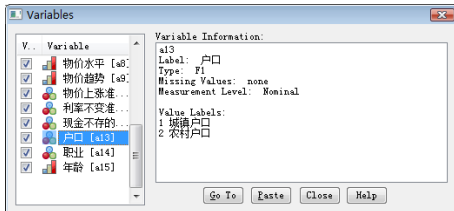
在调查问卷中，涉及到调查人群的性别比例、职称比例，此时可以通过频数分析来得到直观的分析结果。频数分析包含两部分内容：

- 频数、频率表
- 图形，可选择条形图（Bar Chart）、饼图（Pie Chart）、直方图（Histograms）等。可以在直方图上附加正态分布曲线，以便与正态分布比较。



# 频数分析举例

在居民储蓄数据中有户口变量，我们可以对户口变量进行分析，得出调查问卷是否覆盖农村和城市，是否有偏。 如图所示：



# 描述性分析( DESCRIPTIVE STATISTICS)

定距型数据计算描述性统计量，可以得到对定距型变量的一个精确的认识。 基本描述统计量分为三种：

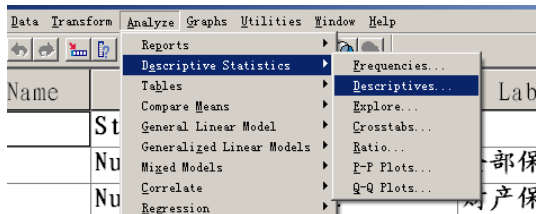
- 描述集中趋势的统计量，如均值（Mean）、中位数（Median）、众数（Mode）、百分位数（Percentile Value）等。
- 描述离散程度的统计量，如样本方差（Variance）、样本标准差（Standard Deviation）、均值标准误差（S.E.Mean）、全距（Range）等。
- 描述分布形态的统计量，如偏度（Skewness）和峰度（Kurtosis）。



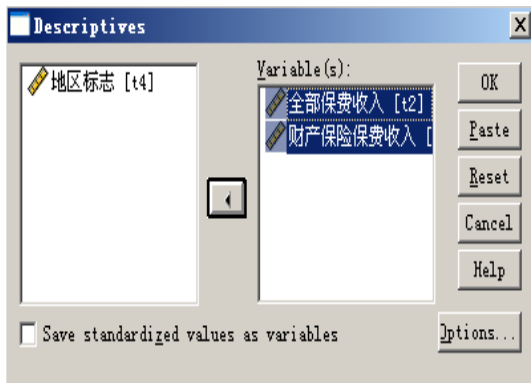
# 描述性分析的基本操作

打开分析窗口

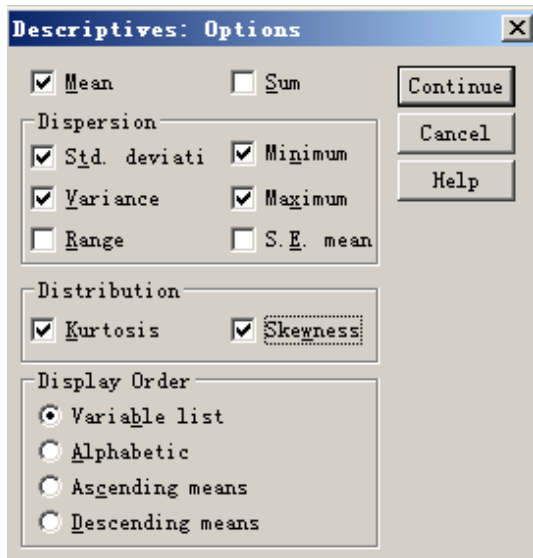
点击 Analysis⇒Descriptive Statistics⇒ Descriptive



弹出窗口，选择待分析的变量到VARIABLE 框中



点击OPTION，选择需要计算的统计量，如样本均值和样本方差



The image shows the 'Descriptives: Options' dialog box in SPSS. It contains several sections for selecting statistical measures:

- Mean and Sum:** ☒ Mean, ☐ Sum
- Dispersion:**
  - ☒ Std. deviation, ☒ Minimum
  - ☒ Variance, ☒ Maximum
  - ☐ Range, ☐ S.E. mean
- Distribution:**
  - ☒ Kurtosis, ☒ Skewness
- Display Order:**
  - ☒ Variable list
  - ☐ Alphabetic
  - ☐ Ascending means
  - ☐ Descending means

Buttons on the right: Continue, Cancel, Help.



常用统计量，其中Kurtosis为峰度，Skewness为偏度，选好后

分析的结果如图所示，得出了指定变量的均值、方差、均方差

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
全部保费收入	36	4152.4347	3267.59230	10677159
财产保险保费收入	36	1703.8014	1364.96325	1863124.7
Valid N (listwise)	36			



# 辛普森悖论

## EXAMPLE

为了研究某种新药对一种疾病的疗效，选择了800名患者作试验，其中400个患者给予新药治疗，另外400个患者给予传统药物治疗来作为对比，实验结果如下

800名患者服药品种与疗效的关系

	疗效		Total
	有效	无效	
新药      新药	200	200	400
	50.0%	50.0%	100.0%
传统药	240	160	400
	60.0%	40.0%	100.0%
Total	440	360	800
	55.0%	45.0%	100.0%

从试验的结果来看，使用新药的有效率为50%，低于传统药物治疗的有效率60%。





## 采用置信度为0.01的 $\chi^2$ 的独立性检验的检验结果

- $H_0$ : 采用哪种药物于疗效是相互独立的
- 可以计算出 $\chi^2 = 8.081$ , 拒绝原假设, 认为新药的疗效低于传统药物的疗效

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.081 <sup>b</sup>	1	.004	.006	.003
Continuity Correction <sup>a</sup>	7.682	1	.006		
Likelihood Ratio	8.095	1	.004		
Fisher's Exact Test					
Linear-by-Linear Association	8.071	1	.004		
N of Valid Cases	800				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 180.00.



# 分成男女两个群体分别进行比较

新药 \* 疗效 Crosstabulation

性别				疗效		Total
				有效	无效	
男	新药	新药	Count	120	180	300
			% within 新药	40.0%	60.0%	100.0%
	传统药	传统药	Count	30	70	100
			% within 新药	30.0%	70.0%	100.0%
	Total	Total	Count	150	250	400
			% within 新药	37.5%	62.5%	100.0%
女	新药	新药	Count	80	20	100
			% within 新药	80.0%	20.0%	100.0%
	传统药	传统药	Count	210	90	300
			% within 新药	70.0%	30.0%	100.0%
	Total	Total	Count	290	110	400
			% within 新药	72.5%	27.5%	100.0%



# $\chi^2$ 检验的结果

Chi-Square Tests

性别		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
男	Pearson Chi-Square	3.200 <sup>b</sup>	1	.074	.075	.046
	Continuity Correction <sup>a</sup>	2.788	1	.095		
	Likelihood Ratio	3.271	1	.071		
	Fisher's Exact Test					
	Linear-by-Linear Association	3.192	1	.074		
	N of Valid Cases	400				
女	Pearson Chi-Square	3.762 <sup>c</sup>	1	.052	.054	.033
	Continuity Correction <sup>a</sup>	3.277	1	.070		
	Likelihood Ratio	3.936	1	.047		
	Fisher's Exact Test					
	Linear-by-Linear Association	3.752	1	.053		
	N of Valid Cases	400				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 37.50.

c. 0 cells (.0%) have expected count less than 5. The minimum expected count is 27.50.



## 上述悖论的解释

可以从全概率公式的角度对上述结果进行解释。



P-P plot[ P-P图]



# MEANS过程简介

means过程可以输出各类群体的某个属性的平均值,可以让分析者对数据或变量一个总的认识

点击Analyze⇒Compare means⇒means..., 如下图

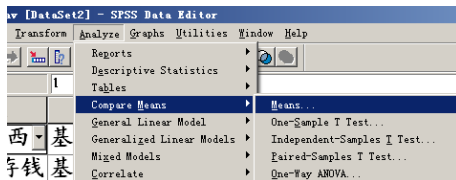
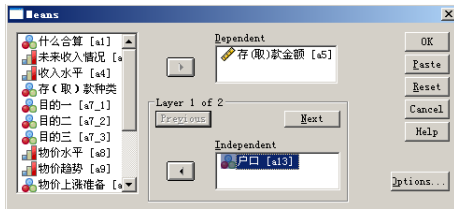


FIG: 均值过程



# 计算不同户口和不同年龄段的一次存款额的平均值



## 第二个层次



选择完毕后，点击Ok即可，另option选项可以选择输出方差分析表(对第一个层次)





# 分析结果

## Report

存(取)款金额

户口	年龄	Mean	N	Std. Deviation
城镇户口	20岁以下	17016.67	3	28568.354
	20~35岁	4682.43	99	8069.112
	35~50岁	5858.64	66	12289.898
	50岁以上	2815.81	32	4427.923
	Total	4956.93	200	9792.515
农村户口	20岁以下	700.00	1	.
	20~35岁	4456.45	47	10395.958
	35~50岁	5032.04	25	19837.091
	50岁以上	977.78	9	987.140
	Total	4204.32	82	13402.425
Total	20岁以下	12937.50	4	24711.515
	20~35岁	4609.68	146	8848.907
	35~50岁	5631.55	91	14634.182
	50岁以上	2412.34	41	3997.900
	Total	4738.09	282	10945.569

FIG: means过程输出结果



# 单样本T检验简介

单样本t检验的目的是检验正态总体 $X \sim N(\mu, \sigma^2)$ 的均值是否为给定值 $\mu_0$ 。它是对总体均值的假设检验，其原假设为

$$H_0: \mu = \mu_0, \longleftrightarrow H_1: \mu \neq \mu_0$$

检验统计量取

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

其中 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 。由数理统计的知识可知，当原假设成立时，上述检验统计量服从参数为 $(n-1)$ 的t分布。SPSS将自动计算t统计量的值 $t_0$ ，并计算概率 $p$ 值 $= P(t(n-1) > t_0)$ 。概率 $p$ 值越小说明t统计量的值越大，当 $p$ 值小于显著性水平 $\alpha$  (通常取 $\alpha = 0.05$ )时应拒绝原假设。



# 打开单样本T检验的主窗口

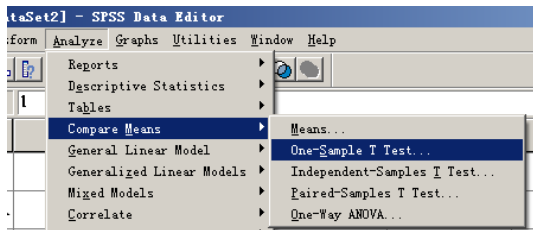


FIG: 单样本t检验



## 检验抽样群体的平均身高与170CM是否有显著差别

在Test框中写入数据170，为原假设，默认为0，点击option弹出下面小窗口，可以输出置信区间，通常选置信水平为0.95

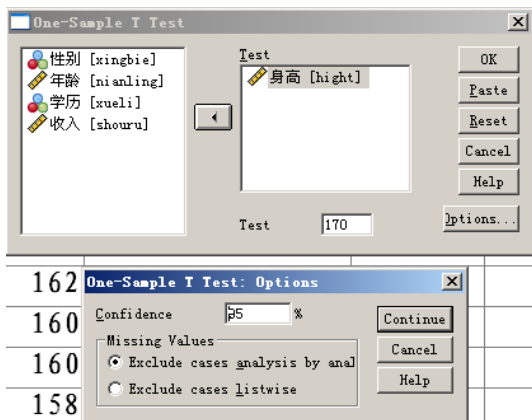


FIG: 单样本t检验



点击CONTINUE，主窗口中点击OK即可输出如下分析结果

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
身高	300	166.64	9.391	.542

One-Sample Test

	Test Value = 170					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
身高	-6.197	299	.000	-3.360	-4.43	-2.29

FIG: 单样本t检验

其中sig即为上述概率  $p$  值，偏小，也可以看出计算出的均值与170差别太大，差值的置信区间不包含0，故拒绝原假设。



## 两独立样本T检验概述

设有两个总体，总体 $X \sim N(\mu_1, \sigma_1^2)$ ，总体 $Y \sim N(\mu_2, \sigma_2^2)$ ，分别来自这两个总体的两组独立样本 $(X_1, X_2, \dots, X_n)$ ， $(Y_1, \dots, Y_m)$ 和相应的样本观察值。做如下检验

$$H_0: \mu_1 = \mu_2 \longleftrightarrow H_1: \mu_1 \neq \mu_2$$

当 $\sigma_1^2, \sigma_2^2$ 未知，但 $\sigma_1^2 = \sigma_2^2$ 时，我们采用如下检验统计量：

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_w \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

其中 $S_w = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ ， $S_1$ 和 $S_2$ 分别为两样本标准差。当原假设成立时， $t \sim t(n+m-2)$ ，SPSS会自动计算相应的t统计量的值和概率p值。

### EXAMPLE

我们可以比较男性和女性的身高是否有显著差别，此时男女分成两个总体，即以性别作为分组变量，比较两组的身高



## 两独立样本T检验的操作，打开主窗口

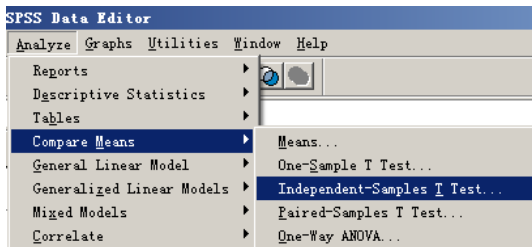


FIG: 两样本t检验的主窗口



## 对不同的性别比较身高是否存在显著的差异

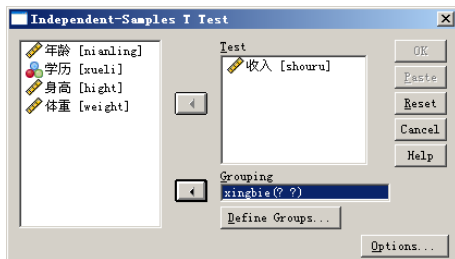


FIG: 两样本t检验

注意分组变量xingbie(??),OK键无法点击,需要定义群组,见下页





点击DEFINE GROUPS按钮，定义不同的群组

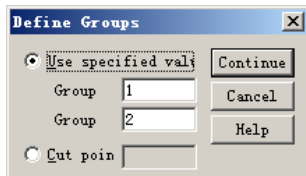


FIG: 定义群组

可以用变量的取值数字作为群组的标示，点击continue，再点击主窗口中的OK键，可得如下输出



## 两样本T检验输出结果

Group Statistics					
	性别	N	Mean	Std. Deviation	Std. Error Mean
身高	男	140	160.99	7.694	.650
	女	160	171.58	7.839	.620

Independent Samples Test									
		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference
									Lower Upper
身高	Equal variances assumed	.345	.557	-11.772	298	.000	-10.588	.899	-12.358 -8.818
	Equal variances not assumed			-11.787	294.081	.000	-10.588	.898	-12.358 -8.820

原假设为身高无显著差别，检验的结果是sig(2-tailed)双边尾概率接近于0，小于0.05,故应拒绝原假设，即认为男女的身高有显著差别。另：男女之间收入的差别可以供同学们练习。



## 两配对样本T检验简介

### EXAMPLE

减肥茶数据中有两个变量，hcg是参与试验的被观察者喝茶前体重，hch为该被观察者喝减肥茶之后的体重， 我们的目的是考察减肥茶是否有效。

一般地，我们抽到的样本 $(X_1, Y_1), \dots, (X_n, Y_n)$ 为n组配对数据，两个数据为同一个对象测出，分别地，我们可以假设 总体 $X \sim N(\mu_1, \sigma_1^2)$ , 总体 $Y \sim N(\mu_2, \sigma_2^2)$ , 我们现在想知道的是 $\mu_1$ 是否等于 $\mu_2$ 。进行如下转换

$$Z_i = X_i - Y_i$$

这样，两配对总体就转化为一个总体 $Z \sim (\mu_1 - \mu_2, \sigma^2)$ 了，于是检验就转化为单总体t检验的情况。SPSS中有专门的Paired-Samples T Test来解决此类检验问题



# 打开PAIRED-SAMPLES T TEST主窗口

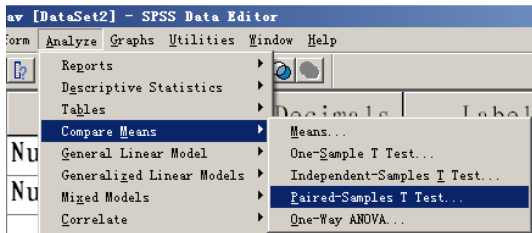


FIG: 两配对样本t检验



# 主窗口

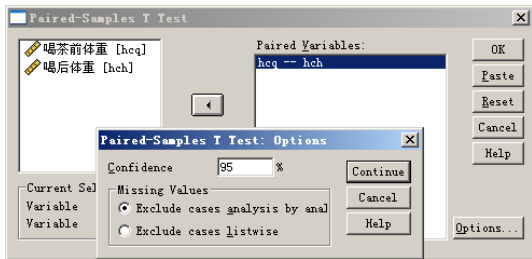


FIG: 两配对样本t检验主窗口

选中两个变量到Paired Variables框中，点击option弹出小窗口，可以给出两均值差的置信区间，点击continue继续， 点击OK可得分析结果



## 两配对样本T检验分析结果

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	喝茶前体重	89.2571	35	5.33767	.90223
	喝后体重	70.0286	35	5.66457	.95749

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	喝茶前体重 & 喝后体重	35	-.052	.768

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	喝茶前 体重 - 喝茶后 体重	19.22857	7.98191	1.34919	16.48669	21.97045	14.252	34	.000

可以看出概率p值接近于0，均值差的置信区间不包含0，故应拒绝原假设，即认为喝茶前后体重有显著差别，减肥茶有效。



# 方差分析的基本思想 I

- 在农业、商业、种植过程中，低投入高产出是人们所预期的，为了实现预定目标，研究人员需要对影响农作物产量的各种因素进行定量的研究，在此基础上制定最佳的种植组合方案，研究过程为
  - ① 找到影响农作物产量的各种因素，如品种、施肥量、地域特征等
  - ② 对不同的品种、不同的施肥量进行比较分析，研究哪个品种的产量高，施肥量为多少合适，品种和施肥量怎么搭配最合适等
  - ③ 在分析的基础上，可以计算出各个种植组合方案的成本和收益，并选择最合理的种植方案。
- 再例如在进行商品的广告宣传时，不同的组合方案所获得的广告效果是不一样的，广告的效果会受到广告形式、地区规模、播出栏目、播放时段、播放频率的影响。人们需要研究在影响广告效果的各种因素中的主要因素，找出最合理的搭配等
- 上述问题的研究都可以通过方差分析来实现



# 方差分析的基本概念

- 在方差分析中，上述问题中的农作物产量、广告效果等称为观测因素或观测变量
- 品种、施肥量、广告形式、播放时段等称为控制因素或控制变量
- 控制变量的不同类别，如品种甲、品种乙等，称为控制变量的不同水平

方差分析正是从观测变量的方差入手，研究诸多控制变量中哪些变量是对观测变量有显著影响的变量。对观测变量有显著影响的各个控制变量的不同水平以及各水平的交互搭配是如何影响观测变量的。





# 方差分析的基本思想 I

方差分析认为导致观测变量变化的因素有两类

- 第一类为控制变量的不同水平产生的影响
- 第二类为随机变量所产生的影响

方差分析认为，如果观测变量值在某控制变量的不同水平中出现了显著波动，则认为该控制变量是影响观测变量的主要因素。反之，如果观测变量值在某控制变量的各个水平中没有出现明显波动，则认为该控制变量没有对观测变量产生重要影响，其数据的波动是抽样误差造成的。然而，如何判断在控制变量不同水平上观测变量的值发生了显著的波动呢？判断的依据是在控制变量的不同水平上，观测变量的分布产生了显著的差别。如果控制变量只有一个，且只有两个水平，我们可以通过前面的两样本t检验进行检验。

方差分析正是通过推断控制变量各水平下观测变量的总体分布是否有显著差异来实现统计分析的。



# 方差分析的基本思想 II

方差分析对观测变量各总体分布的两个基本假设前提:

- ① 观测变量各总体应服从正态分布
- ② 观测变量各总体的方差应相同

基于上述假设，方差分析对各总体分布是否有显著差异的推断就转化为各总体均值是否有显著差异的推断。



# 方差分析的分类

## 方差分析的分类

- 根据控制变量的个数，可以将方差分析分为单因素方差分析、多因素方差分析以及协方差分析
- 观测变量为一个以上的方差分析称为多元方差分析

关于方差分析，本课程主要介绍单因素方差分析。



# 单因素方差分析的基本思想 I

单因素方差分析用来研究一个控制变量的不同水平是否对观测变量产生了显著的影响。下述问题均可以通过单因素方差分析解决:

- 例如分析不同施肥量是否给农作物的产量带来显著性影响
- 考察学历对工资收入的影响等

## ① 第一步, 明确观测变量和控制变量

如上述例子中观测变量分别是农作物产量、工资收入, 控制变量为施肥量、学历。

## ② 第二步, 剖析观测变量的方差 (离差平方和)

方差分析认为观测变量的变动受到控制变量的水平和随机因素的共同影响, 我们不妨设观测变量为 $Y$ , 控制因素记为 $X$ ,  $X$ 有 $k$ 个水平, 分别记为 $x_1, x_2, \dots, x_k$ , 在 $X$ 的第 $i$ 个水平 $x_i$ 处, 得到了 $Y$ 的 $n_i$ 次观测, 记为 $y_{ij}, j = 1, \dots, n_i$ , 那么我们可以得到

$$y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \dots, k, j = 1, \dots, n_i$$



## 单因素方差分析的基本思想 II

其中 $\mu_i$ 为观测变量在第 $i$ 个水平下的理论值， $\varepsilon_{ij}$ 为抽样误差，为服从正态分布的随机变量。

如果令

$$\mu = \frac{1}{k} \sum_{i=1}^k \mu_i$$

其中 $\mu$ 为观测变量总的理论值，且令

$$a_i = \mu_i - \mu$$

其中 $a_i$ 为控制变量 $X$ 的水平 $x_i$ 对试验结果产生的附加影响，称为水平 $x_i$ 对观测变量产生的效应，有 $\sum_{i=1}^k a_i = 0$ 。  
这样我们就可以得到

$$y_{ij} = \mu + a_i + \varepsilon_{ij}, i = 1, \dots, k, j = 1, \dots, n_i$$

上式就是单因素方差分析的数学模型，如果控制变量 $X$ 对观测变量没有影响，则各水平的效应 $a_i$ 应该全部为0，否则就



## 单因素方差分析的基本思想 III

不全为0。单因素方差分析正是要对控制变量的所有效应是否同时为0进行推断。

上述模型中 $\mu$ 的无偏估计为

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij},$$

其中 $n = \sum_{i=1}^k n_i$ 为总的样本数,  $\mu_i$ 的无偏估计为第*i*组观测变量的平均值,

$$\hat{\mu}_i = \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij},$$

由此可以得到观测变量的总的离差平方和为

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2,$$



## 单因素方差分析的基本思想 IV

方差分析的做法是将上述离差平方和分解为组间平方和与组内平方和(按照因素的不同水平对 $Y$ 的观测进行分组),

$$SST = SSA + SSE,$$

其中

$$SSA = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

可见: 组间离差平方和 $SSA$ 是各水平均值和总均值离差的平方和, 反映了控制变量不同水平对观测变量的影响。

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,$$

组内平方和 $SSE$ 是每个样本数据与本组均值离差的平方和, 反映了随机因素对观测变量的影响。



## 单因素方差分析的基本思想 V

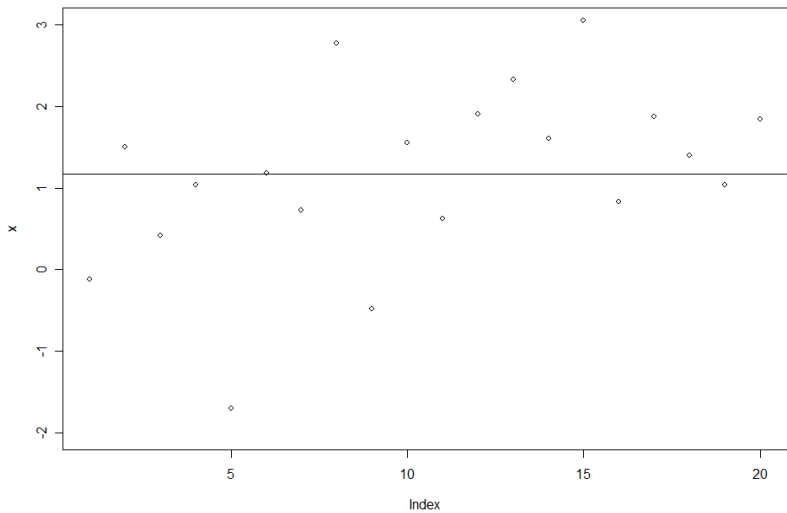
- ③ 第三步，比较观测变量总离差平方和中两个部分的比例。显而易见，在总的离差平方和中，如果组间离差平方和(SSA)占的比例较大，则说明观测变量的变动 主要是由控制变量引起的，可以由控制变量来解释观测变量的变化；如果SSE占的比例较大，说明观测变量的 变动不是由控制变量的变化引起，更可能是由随机因素的影响造成的。鉴于随机因素可能与测量单位有关，故而不恒定，我们取下述统计量

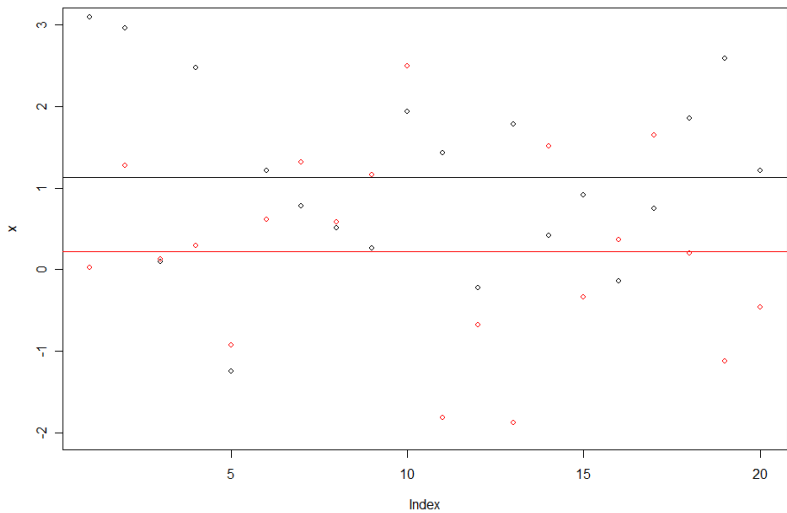
$$F = \frac{SSA/(k-1)}{SSE/(n-k)}$$

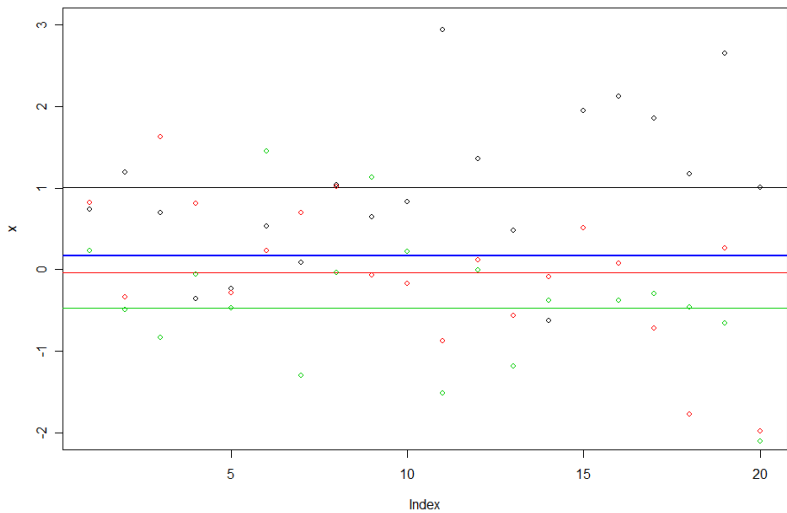
其中 $n$ 为总的样本数， $k-1$ 和 $n-k$ 分别为SSA和SSE的自由度。











# 单因素方差分析的步骤 I

方差分析问题属于统计推断中的假设检验问题，其基本步骤与假设检验完全一致

## ① 提出零假设

单因素方差分析的零假设 $H_0$ 是：控制变量不同水平下观测变量各总体的均值无显著差别，即各水平的效应同时为0，记为：

$$\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$

## ② 选择检验统计量

方差分析采用的统计量是F统计量：

$$F = \frac{SSA/(k-1)}{SSE/(n-k)} = \frac{MSA}{MSE}$$

MSA为平均组间平方和，MSE为平均组内平方和，其目的是为了消除水平数和样本数对分析产生的影响。F统计量服从自由度为  $(k-1, n-k)$  的F分布。



## 单因素方差分析的步骤 II

### ③ 计算检验统计量的观测值和概率p值

该步的目的是计算检验统计量的观测值和相应的概率p值。

概率p值 =  $P\{F(k-1, n-k)r.v. > F_0\}$ , 其中  $F_0$  为  $F$  统计量的观察值。 不难理解, 如果控制变量对观测变量产生了显著的影响, 观测变量的总的离差平方和中SSA所占的比例对于SSE必然较大,  $F$  值将偏大,  $p$  值偏小; 反之, 如果控制变量没有对观测变量产生显著影响, 总的离差平方和 SST 应归结于误差造成的影响,  $F$  值将偏小,  $p$  值就偏大

### ④ 给出显著性水平 $\alpha$ , 做出决策

给出显著性水平, 与检验统计量的概率p值相比较。如果

- 概率p值小于给定的 $\alpha$ 值, 则应拒绝零假设, 认为控制变量不同水平下观测变量各总体的均值存在显著差异;
- 概率p值大于给定的 $\alpha$ 值, 则不应拒绝零假设, 认为控制变量不同水平下观测变量各总体的均值无显著差异。



# 单因素方差分析的基本操作

## 单因素方差分析操作举例

广告形式对销售额的单因素方差分析



# 方差齐性检验

方差齐性检验是对控制变量的不同水平下各观测变量总体方差是否相等进行分析。

SPSS单因素方差分析中，方差齐性检验采用了方差同质性(Homogeneity of Variance)的检验方法，其零假设是各水平下观测变量总体方差无显著差别，实现思路同两独立样本t检验中的方差检验。



# 多重比较检验

单因素方差分析的基本分析只能判断控制变量是否对观测变量产生了显著性影响。如果确实影响显著，还需要进一步确定，控制变量的不同水平对观测变量的影响程度如何，其中哪个水平的作用明显不同于其他水平，哪个水平的作用不显著等。。。。，如广告的形式有 报纸、广播、体验、宣传品，到底哪种形式更能促进销售呢，两种形式之间的差别如何？？？显然，我们可以通过两独立样本t检验来解决，对各个水平下观测变量总体均值进行两两逐对检验。这样的比较需要进行多次。但是这样的比较会使犯第一类错误的概率增加，而多重比较检验正是解决该问题的一类方法。

多重比较问题涉及到较多的统计学知识，理论比较难，但是操作并不困难，困难的是其统计解释，希望同学们有所侧重





# 多重比较检验的几种方法 I

- LSD方法

也称为最小显著性差异(Least Significant Difference)方法。

采用t统计量，定义为

$$t = \frac{(\bar{x}_i - \bar{x}_j) - (\mu_i - \mu_j)}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}}$$

其中MSE为观测变量的组内方差。正因为如此，不同于前述两独立样本t检验。这里t统计量服从自由度为n-k的t分布。LSD方法适用于各总体方差相等的情况，但没有有效地控制第一类错误概率。



## 多重比较检验的几种方法 II

- Bonferroni方法

Bonferroni方法与LSD基本相同，不同的是Bonferroni 方法对犯第一类错误概率进行了控制。在每一次的两两组的检验中，将显著性水平由  $\alpha$  变成  $\alpha/N$ ，其中N为两两检验的总次数，从而从总体上控制了犯第一类错误概率。两总体均值差的置信区间为

$$(\bar{x}_i - \bar{x}_j) \pm t_{\frac{\alpha}{2N}}(n - k) \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$



## 多重比较检验的其他方法

- Tukey方法，方法适用于各水平观测值个数相等的情况，即 $n_i \equiv r, i = 1, \dots, k$ ,其他类似于Boferroni方法
- Scheffe方法，与Tukey方法相比不够灵敏
- S-N-K方法,是一种有效划分相似性自己的方法，该方法只适用于各水平观测值个数相等的情况



## 其他检验 I

- 先验对比检验

在多重比较检验中，如果发现某些水平与另一些水平的均值差距显著，比如有5个水平，其中 $\bar{x}_1, \bar{x}_2, \bar{x}_3$ 与 $\bar{x}_4, \bar{x}_5$ 有显著差异，就可以进一步比较这两组总的均值是否存在显著差异，即 $\frac{1}{3}(\bar{x}_1 + \bar{x}_2 + \bar{x}_3)$ 和 $\frac{1}{2}(\bar{x}_4 + \bar{x}_5)$ 是否有显著差别，这种比较分析实际上是对各均值线性组合结果的分析。即如果令 $c_1 = c_2 = c_3 = \frac{1}{3}, c_4 = c_5 = -\frac{1}{2}$ ，且 $\sum_{i=1}^5 c_i = 0$ ，则应推断

$$\sum_{i=1}^5 c_i \bar{x}_i \text{ 是否显著为 } 0.$$

这种事先指定各均值的系数，然后对其线性组合进行检验的分析办法称为先验对比检验，通过先验对比检验能够更精确地掌握各水平间或各相似性子集间均值的差异程度。



## 其他检验 II

- 趋势检验(Contrast) 当控制变量为定序变量是，趋势检验能够分析随着空时变量水平的变化，观测变量值变化的总体趋势是怎样的。通过趋势检验，能够帮助我们 把握控制变量不同水平对观测变量总体作用的程度。



# 单因素方差分析进一步分析的操作

单因素方差分析的进一步分析的选项安排在下图的三个按钮选项中

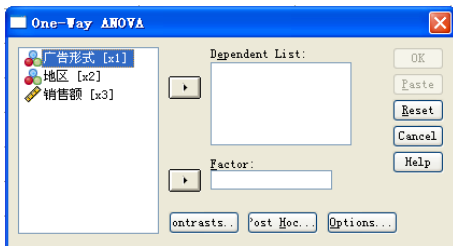
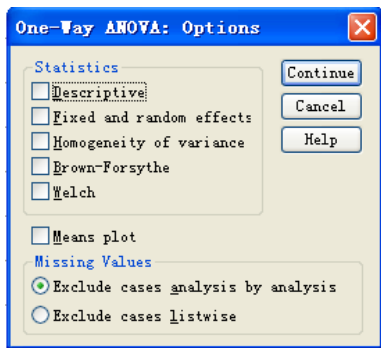


FIG: 单因素方差分析进一步分析的操作

Contrasts, Post Hoc, Option三个按钮的选项中，提供了进一步分析的工具。



# OPTION 选项

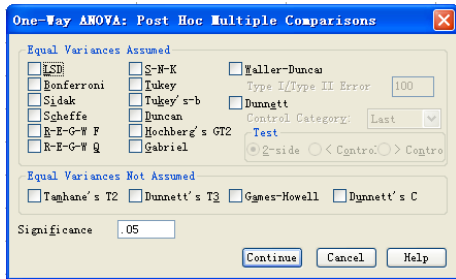


Option选项用来对方差分析的前提条件进行检验，并可输出其他相关统计量和对缺失数据进行处理。Option的窗口选项见图形,所示的选项中，Homogeneity of Variance test选项实现方差齐性检验，Means Plot 选项输出各水平下观测变量均值的折线图。Descriptive选项输出观测变量的基本描述统计量，Missing

Values框中提供了两种缺失数据的处理方式。



# Post Hoc选项



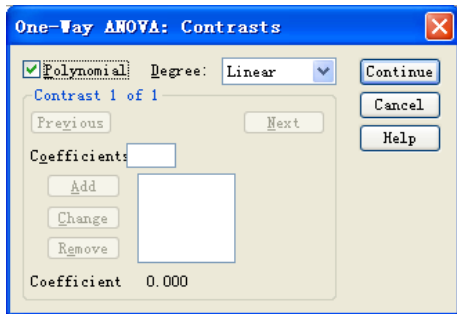
Post Hoc 选项用来实现多重比较检验。如窗口所示，提供了18种多重比较检验的方法，其中Equal Variance Assumed 框中的方法适用于各水平方差齐性的情况，Equal Variance Not Assumed 中的方法适

用于各水平方差不齐的情况。在方差分析中，由于前提所限，在应用中多采用方差齐性的情况。





## CONTRAST 选项



次类推.....。

Contrast选项用来实现对比检验和趋势检验.如果进行趋势检验,则应选择Polynomial选项,然后再后面的下拉框中选择趋势检验的方法。其中Linear表示线性趋势检验,Quadratic表示进行二次多项式的,cubic表示三次多项式,依



## Part VIII

### 假设检验的理论补充



# 假设检验的思想

在现实的均值比较和列联表分析中，都会出现假设检验问题

## EXAMPLE (两独立总体比较问题)

两总体  $X \sim N(\mu_1, \sigma_1^2)$ ,  $Y \sim N(\mu_2, \sigma_2^2)$ , 其中  $\sigma_1^2, \sigma_2^2$  未知, 试考虑如下问题

- 均值是否相等:  $H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \mu_1 \neq \mu_2$
- 方差齐性检验:  $H_0: \sigma_1 = \sigma_2 \leftrightarrow H_1: \sigma_1 \neq \sigma_2$

## EXAMPLE (独立性检验)

两个变量: 人种, 高血压, 试问如下问题

$H_0$ : 患高血压的概率与人种无关  $\leftrightarrow H_1$ : 某些人种更容易患高血压



## 如何分析，选取统计量，度量两个假设之间的差距

### EXAMPLE

两独立总体的比较问题 可以分别选取下述统计量

- $t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ , 在方差相同但未知的情况,
- $F = \frac{S_1^2}{S_2^2}$ , 涉及的变量可以由样本计算得出

### EXAMPLE

独立性检验 取统计量:  $K = \sum_{i,j} \frac{(n_{ij} - n \times \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j})^2}{n \times \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j}}$ , 其中的  $\hat{p}_{i\cdot}, \hat{p}_{\cdot j}$  可以有样本计算得出。

根据上述统计量，进而计算出统计量的值后如何说明检验是否成立呢？我们需要知道统计量的零分布



# 统计量的零分布，原假设条件下的分布

根据数理统计的知识，我们可以得到如下结果,在上述三个假设分别成立的时候，我们可以得到

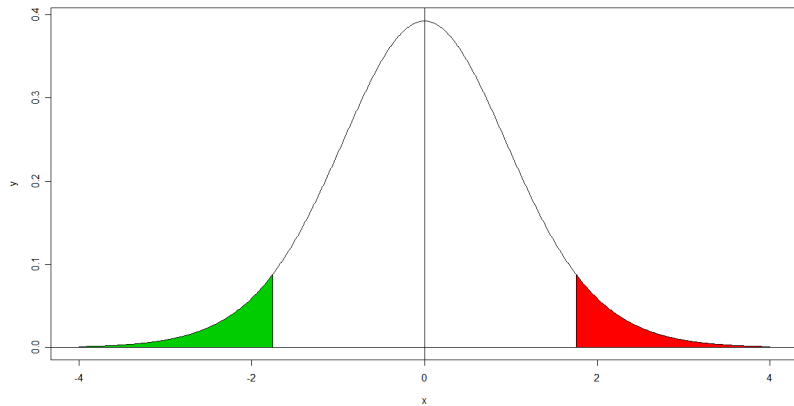
- $\mu = \mu_2$ 时， $t \sim t(n_1 + n_2 - 2)$
- $\sigma_1 = \sigma_2$ 时， $F \sim F(n_1 - 1, n_2 - 1)$
- 人种与高血压独立时， $K \sim \chi^2((a - 1) \times (b - 1))$

并且我们可以根据样本数据算出上述统计量的值，偏大吗，是不是不正常？回答此问题，我们需要尾概率（也常称为p值，显著性水平(Sig...)）

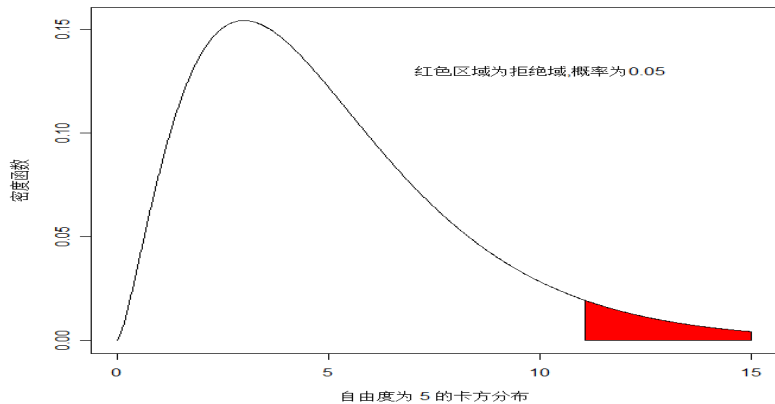
- $P(|t| > t_{value})$ ，其中 $t_{value}$ 是根据样本计算出来的，这是双尾概率(two tailed)
- $P(F > F_{value})$ ，单尾概率
- $P(K > \chi^2_{value})$ ，单尾概率



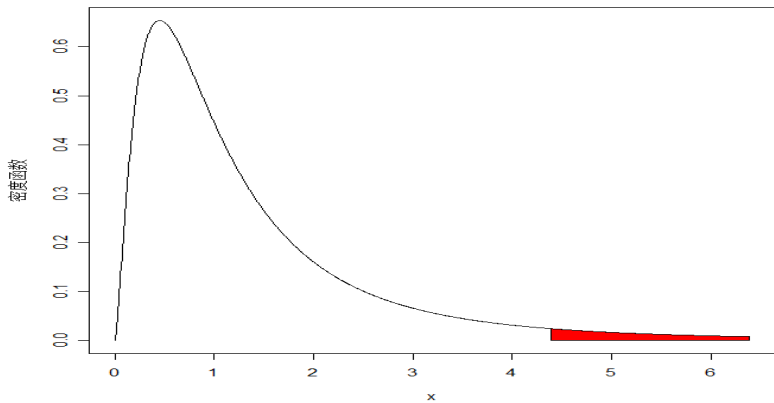
t检验的拒绝域



### 卡方拟合优度检验的拒绝域



方差齐性检验的拒绝域





# 相关分析与回归分析概述

相关分析和回归分析都是分析客观事物之间关系的数量分析方法，明确客观事物之间的关系对理解相关分析和回归分析是极为重要的。客观事物之间的关系大致可以分为两类

- 函数关系
- 统计关系

相关分析和回归分析是用来分析事物之间统计关系的方法



# 函数关系和统计关系

**函数关系** 函数关系是两事物之间的一一对应关系，当一个变量取定后，我们可以 通过函数得到另一个变量的值，如销售额和销售量，如果单价给定，销售额和销售量之间便建立了对应关系 $y = px$ 。客观世界中类似的关系很多，都是确定的。

**统计关系** 统计关系指的是两事物之间的一种非一一对应的关系，即如果一个变量 $x$ 确定，另一个变量 $y$ 的取值无法根据 特定的函数得到。如家庭收入和支出之间的关系，子女身高和父母身高的关系。统计关系可以细分为线性相关和非线性相关。而线性相关还可以再细分为正线性相关和负线性相关等等。。。

**相关分析和回归分析** 提供了测度事物间统计关系的工具。



# 相关分析的基本方法

图形方法 散点图,可以直观的看出两个变量之间的关系

数值方法 相关系数矩阵

- Pearson相关系数(对数值型变量)
- Spearman等级相关系数 (对属性变量)
- Kendall  $\tau$ 相关系数(非参数检验方法,适用于定序变量)



## 散点图的含义

绘制散点图是相关分析过程中极为常用且非常直观的分析方式。将数据以 点的形式画在直角平面上。通过观察散点图能够直观地发现变量间的关系 以及他们的强弱程度和数据对的可能走向。下面几个散点图的形状可以得到变量之间关系。



# 弱相关

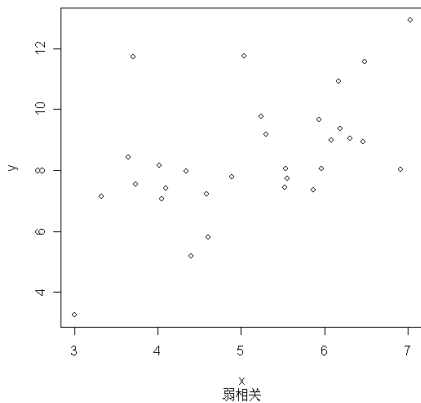
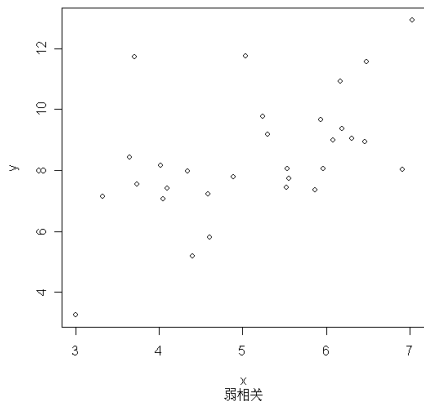


FIG: 弱相关

# 相关分析散点图一

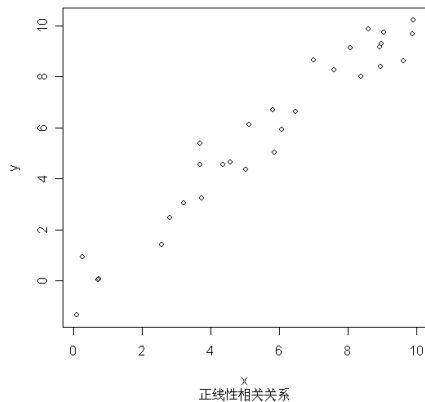


## 弱相关

两个变量之间存在一定的相关性，但是不显著，此时我们称之为弱相关



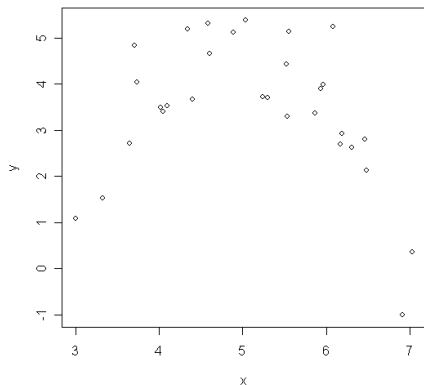
# 强相关



从图形中我们可以看出，两个变量之间存在明显的正线性相关关系



# 非线性相关



从图形中我们可以看出，两个变量之间存在曲线相关关系，但是检验线性相关系数和0非常接近，此时不能称二者不相关，只能成为非线性相关





# 散点图的基本操作

- ① 选择菜单Graphs→Legacy Dialogs →scatter/dot
- ② 选择散点图的类型，SPSS提供了5种类型的散点图。
- ③ 根据所选择的散点图的类型，点击Define 进行具体定义

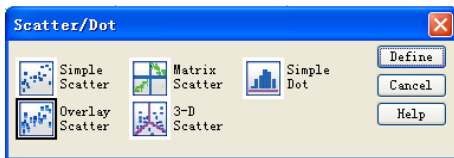


FIG: 散点图类型



# 散点图的分类

- ❶ 简单散点图(simple) 表示一对变量关系的散点图，set Markers by 选项可以在一张图上分别以不同颜色绘制散点图
- ❷ 重叠散点图(Overplay)表示多对变量之间统计关系的散点图
- ❸ 矩阵散点图(Matrix)以方形矩阵的形式在多个坐标轴上分别显示多对变量间的统计关系。
- ❹ 三维散点图(3D scatter) 以立体图的形式展现三对变量间的统计关系
- ❺ 注：
  - 分组变量选在 Set Markers by 框中，对该变量不同的取值，点的形状和颜色可以发生变化
  - Label Cases by 框中为标记变量，则将标记变量的取值标注在散点图中点的旁边。



# 相关系数

散点图可以直观地展现变量之间的统计关系，但并不精确，相关系数以数值的方式精确地反映两个变量之间关系的强弱程度。

计算相关系数 $R$ ，相关系数的计算方法有

PEARSON相关系数 用来度量定矩型变量之间的线性相关关系。

SPEARMAN等级相关系数 用来度量定序变量之间的线性相关关系

KENDALL  $\tau$ 相关系数 采用非参数检验方法用来度量定序变量间的线性相关关系。

$r$ 的取值范围都是一样的， $[-1, 1]$ ，如果 $|r|$ 接近于1，说明两者之间具有较强的线性关系，计算出相关系数之后，需要对两个总体间是否存在显著的线性关系进行推断



# PEARSON简单相关系数

数学定义为:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

其中n为样本数,  $x_i, y_i, i = 1, \dots, n$ 分别为两个变量的变量值, 简单相关系数具有如下特点

- ① x和y 是对等的, x和y的相关系数等于y和x的相关系数
- ② 可以看出相关系数都是对变量进行标准化之后的计算, 因而简单相关系数是无量纲的。
- ③ 对x和y进行线性变化之后可能改变其相关系数的符号, 但不会改变相关系数的值。
- ④ 相关系数能够度量变量间的线性关系, 但不能度量非线性关系



# PEARSON相关系数的检验统计量

一般地，当 $|r| > 0.8$ 时，我们认为两变量之间具有较强的线性关系，当 $|r| < 0.3$ 时，表示两变量之间的线性关系较弱。如何精确地判断呢？我们一般采用如下的检验统计量

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

服从自由度为 $n-2$ 的 $t$ 分布，



# SPEARMAN等级相关系数

Spearman相关系数用来度量定序变量之间的线性相关关系。该系数的设计思想与Pearson相关系数完全相同，然而在计算Spearman相关系数的时候，由于数据是非定矩的，不能直接采用原始数据，而是利用数据的秩，用量变量的秩( $U_i, V_i$ )代替( $x_i, y_i$ )计算公式中

$$r = \frac{\sum_{i=1}^n (U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\sum_{i=1}^n (U_i - \bar{U})^2 \sum_{i=1}^n (V_i - \bar{V})^2}},$$

可以简化为

$$r = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)},$$

其中  $D_i = U_i - V_i$ .



# SPEARMAN相关系数的思想和检验

可见SPearman相关系数体现了如下思想

- 如果两变量的正相关性较强，他们的秩的变化具有同步性，于是 $\sum_{i=1}^n D_i^2$ 的值较小，从而 $r$ 趋近于1
- 当两个变量无安全正线性相关时， $U_i = V_i$ ，此时 $\sum_{i=1}^n D_i^2$ 的值最小，为0，此时 $r=1$
- 如果两变量之间的正线性关系较弱， $\sum_{i=1}^n D_i^2$ 的值偏大， $r$ 接近于0

小样本下，零假设成立时，Spearman等级相关系数服从Spearman分布，在大样本情况下，Spearman等级相关系数的检验统计量为

$$Z = r\sqrt{n-1},$$

渐进服从标准正态分布。SPSS将自动计算Spearman系数， $Z$ 统计量的值和相应的 $p$ 值( $p$ -value)。



## KENDALL $\tau$ 系数

Kendall  $\tau$ 系数采用非参数检验方法来度量定序变量间的线性相关关系，利用变量秩计算一致对(同序对)数目(U) 和非一致对(异序对)数目(V).Kendall  $\tau$ 系数的数学定义为：

$$\tau = (U - V) \frac{2}{n(n-1)},$$

小样本情况下，Kendall  $\tau$ 服从Kendall分布，大样本情况下采用Z检验统计量

$$Z = \tau \sqrt{\frac{9n(n-1)}{2(2n+5)}},$$

渐进服从标准正态分布。SPSS自动计算Kendall  $\tau$ 系数、Z检验统计量的观测值和相应的 概率p值。





# 计算相关系数的基本操作

- ① 选择菜单Analysis→Correlate→Bivariate，弹出对话框
- ② 把参与计算相关系数的变量选到Variables框中
- ③ 在Correlation Coefficients 框中选择计算哪种相关系数
- ④ 在test of Significance 框中选择输出相关系数检验的双尾概率或单尾概率
- ⑤ 选中Flag significance correlations 选项，输出星号标记，以标明变量间的相关性是否显著，不选中则不输出星号标记



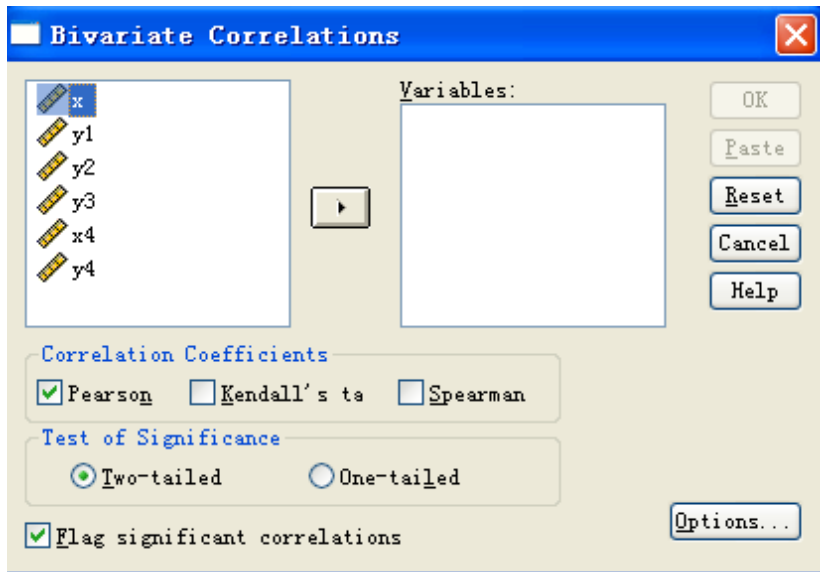


FIG: 相关系数

# 偏相关分析和偏相关系数

相关分析中研究两事物之间的线性相关的强弱，但是就相关系数而言，未必是两事物之间的现行相关强弱的真实体现，例如：

- 研究商品的需求量和价格、消费者收入之间的线性关系时，需求量和价格的相关关系实际上还包含了 消费者收入对商品需求量的影响，同时收入对价格也会产生影响，并通过价格变动传递到对商品需求量的影响中。
- 在这种情况下，单纯利用相关系数来评价变量间的相关性显然是不准确的， 而需要提出其他相关因素影响的条件下计算变量间的相关。
- 偏相关分析的意义就在于此。偏相关分析也称为净相关分析，它在控制其他变量的线性影响的条件下分析两个变量间的相关，所采用的工具是偏相关系数。当控制变量个数为1时，称为1阶偏相关，控制变量个数为2时，称为2阶偏相关。



## 偏相关系数

计算样本的偏相关系数,反映两变量间净相关的程度强弱。在分析变量 $x_1$ 和 $y$ 之间的净相关时,当控制了 $x_2$ 的线性作用后, $x_1$ 和 $y$ 之间的一阶偏相关定义为

$$r_{y1,2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}$$

其中,  $r_{y1}, r_{y2}, r_{12}$ 分别表示 $y$ 和 $x_1$ 的相关系数,  $y$ 和 $x_2$ 的相关系数,  $x_1$ 和 $x_2$ 的相关系数。



# 对样本来自的两总体是否存在显著的净相关进行推断

## 统计推断过程

- ① 提出零假设 $H_0$ ,即两总体的偏相关系数与0无显著差别
- ② 选择检验统计量, 偏相关分析的检验统计量为t检验统计量, 定义为

$$t = r \sqrt{\frac{n - q - 2}{1 - r^2}}$$

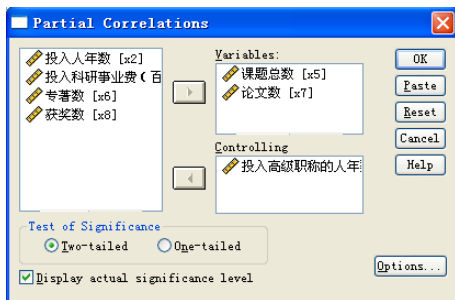
其中 $r$ 为偏相关系数值,  $n$ 为样本数,  $q$ 为阶数。t统计量服从自由度为 $n-q-2$ 的t分布。

- ③ 计算检验统计量的观测值, 和对应的p值
- ④ 决策, 如果检验统计量的p值小于给定的显著性水平 $\alpha$ , 应拒绝原假设。反之接受原假设。



# 偏相关分析的基本操作

- 选择菜单Analyze→Correlate→Partial，弹出对话框



- 报参与分析的变量和控制变量选入对应的对话框
- 选择对立假设单尾或是双尾
- option 中选中Zero-order Correlations 表示输出零阶偏相关系数

SPSS将自动进行偏相关分析和统计检验，并将结果显示到输出窗口(Out Put)中



# 距离分析的基本概念

举例分析(Distances)是对观测变量之间相似或不相似程度的一种测度，是计算一对变量之间或一对观测变量之间的广义的距离。这些相似性或距离测度可以用于其他分析过程，如因子分析，聚类分析等。

在距离分析过程中，主要利用变量间的相似性测度(Similarities)和不相似性测度(Dissimilarities)度量两者之间的关系。



# 相似性测度

两变量之间的相似性，可以用相关系数等度量，

- 针对定距型数据有pearson相关系数和夹角余弦距离等
- 对二值变量的相似性测度主要包括简单匹配系数(Simple matching), Jaccard相似性指数，Hamann相似性测度等20余种
- 另外，相似性测度可以用于因子分析、聚类分析等模块





# 不相似性测度

- 对定距型变量间距离描述的统计量，主要有欧式距离，平方欧式距离，契比雪夫(Chebychev)距离、Block距离、Minkowski距离
- 对于定序型变量之间距离的描述，主要有平方不相似测度(Chi-Square measure)和Phi-Square不相似测度
- 对于二值变量之间的距离描述，主要有欧式距离，平方欧式距离，Lane and Williams不相似测度等



# 距离分析的SPSS操作 I

- ① 打开主对话框，选择Analysis⇒Correlate⇒Distance， 打开Distance(距离分析)主对话框，弹出如下图型



FIG: "Distance(距离分析)" 主对话框



## 距离分析的SPSS操作 II

### ② 选择距离分析变量和分析方式

从左侧的变量列表中选择变量，单击右向箭头按钮，将其移动到“Variable”(变量)列表框，选择个案标识变量到“Label Case”(个案标识)列表框，增加结果的可读性。

在“Compute Distance”选项组中选择计算距离的对象

- Between Cases 个案间距离，计算个案之间的距离。
- Between Variables 变量距离，计算变量之间的距离

在“Measure”选项组中选择计算距离测度的类型，包括如下选项：

- Similarity Measure 相似性测度，数值越大，表示距离越近
- Dissimilarity Measure 不相似性测度，数值越大，说明距离越远。

### ③ 执行操作，单击“OK”，执行距离分析的操作。



# 回归分析概述

回归分析是一种应用极为广泛的数量分析方法，用于分析事物之间的统计关系，侧重考察变量之间的数量变化规律，并通过回归方程的形式描述和反映这种关系，帮助人们准确把握变量受其他一个或多个变量影响的程度，进而为控制和预测提供科学依据。如下图,中间穿过散列点的线称为**回归线**

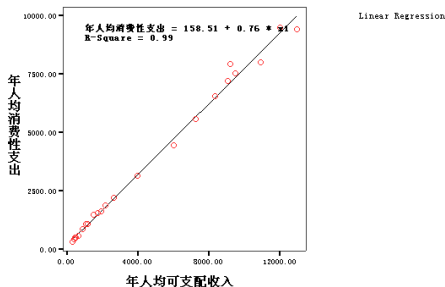


FIG: 回归分析示意图



# 回归线和回归模型

## 得到回归线的方法

- 局部平均
- 函数拟合。假定变量间的关系符合某个函数，这样只需要顾及函数的参数就可以了。

## 回归分析的一般步骤

- ① 确定回归方程中的解释变量（即自变量）和被解释变量（即因变量）
- ② 确定回归模型。回归模型分为线性回归模型和非线性回归模型两类。
- ③ 建立回归方程
- ④ 对回归方程进行各种检验
- ⑤ 利用回归方程进行预测



# 线性回归模型

观察被解释变量 $y$ 和一个或多个解释变量 $x_i$ 的散点图，当发现 $y$ 与 $x_i$ 之间呈现显著的线性关系时，则应采用线性回归分析的方法，建立 $y$ 关于 $x_i$ 的线性回归模型，在线性回归分析中，根据模型中解释变量的个数，线性回归模型分为

- 一元线性回归模型,只有一个解释变量
- 多元线性回归模型,有多个解释变量



# 一元线性回归模型 I

一元线性回归模型是指只有一个解释变量的线性回归模型，用于揭示被解释变量和解释变量之间的线性关系。一元线性回归的数学模型是：

$$y = \beta_0 + \beta_1 x + \epsilon \quad (1)$$

上式表示被解释变量 $y$ 的变化可由两部分解释

- 由于解释变量 $x$ 的变化引起的 $y$ 的线性变化部分，即 $y = \beta_0 + \beta_1 x$
- 由于随机因素引起的 $y$ 的变化部分，即 $\epsilon$

$\beta_0, \beta_1$ 为模型中的未知参数，它们分别被称为回归常数和回归系数， $\epsilon$ 称为随机误差，通常它是一个随机变量，并且满足

$$\begin{cases} E(\epsilon) = 0; \\ \text{var}(\epsilon) = \sigma^2 \end{cases}$$

对方程(1)两边同时求期望可得：

$$E(y) = \beta_0 + \beta_1 x$$



## 一元线性回归模型 II

上式称为一元线性回归方程，表明 $x$ 和 $y$ 之间的统计关系是在平均意义下表述的。对回归方程(2)中未知参数 $\beta_0, \beta_1$ 进行估计是一元线性回归分析的核心任务之一。由于参数估计是基于样本数据的，由此得到的参数只是参数真值的估计值，记为 $\hat{\beta}_0, \hat{\beta}_1$ ，于是有

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (3)$$

上式常称为一元线性经验回归方程。

### 几何意义

从几何上讲，一元线性经验回归方程就是二维平面上的一条直线，即回归直线。其中 $\hat{\beta}_0$ 是回归直线在纵轴上的截距， $\hat{\beta}_1$ 是回归直线的斜率，它表示解释变量 $x$ 每变动一个单位所引起的被解释变量 $y$ 的平均变动单位。





# 多元线性回归模型 I

多元线性回归模型是指有多个解释变量的线性回归模型，用来揭示被解释变量与其他多个解释变量之间的线性关系。多元线性回归的数学模型是：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon \quad (4)$$

上式(4)是一个p元线性回归模型，其中有p个解释变量。



## 多元线性回归模型(续)

p元线性回归方程为

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (5)$$

p元线性经验回归方程为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p \quad (6)$$

从几何上讲，它是p维空间上的一个超平面，即回归平面。



# 回归参数估计的最小二乘法 I

线性回归方程确定后的任务是利用已经收集到的数据，根据一定的统计拟合准则，对方程中的各个参数进行估计。

## 普通最小二乘法的基本思想

普通最小二乘估计的基本出发点就是使每个样本点 $(x_i, y_i)$ 到回归线的对应点 $(x_i, E(y_i))$ 在竖直方向上的偏差距离的总和最小。应如何定义这个偏差距离呢？普通最小二乘法将这个偏差距离定义为离差的二次方，即 $(y_i - E(y_i))^2$ ，于是竖直方向上偏差距离的总和就是离差平方和

### ❶ 对于一元线性回归模型

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - E(y_i))^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (7)$$



## 回归参数估计的最小二乘法 II

最小二乘法就是寻找参数 $\beta_0, \beta_1$ 的估计值 $\hat{\beta}_0, \hat{\beta}_1$ , 使上式达到最小。即

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

### ② 对于多元线性回归模型

$$Q(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_1 - \dots - \beta_p x_p)^2$$

最小二乘法是寻找参数 $\beta_0, \beta_1, \dots, \beta_p$ 的估计值 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$  使得上式达到最小, 即

$$Q(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \min_{\beta_0, \beta_1, \dots, \beta_p} Q(\beta_0, \beta_1, \dots, \beta_p)$$

在使用SPSS分析时, 系统会自动完成参数的估计并给出最终的估计值。



# 回归分析的简单基本操作及参数的估计 I

SPSS中一元线性回归和多元线性回归的功能菜单是集成在一起的，下面给出操作步骤和对话框介绍

- 1 打开主对话框，选择"Analysis"  $\Rightarrow$  "Regression"  $\Rightarrow$  "Linear"，打开"Linear Regression"对话框,如图所示

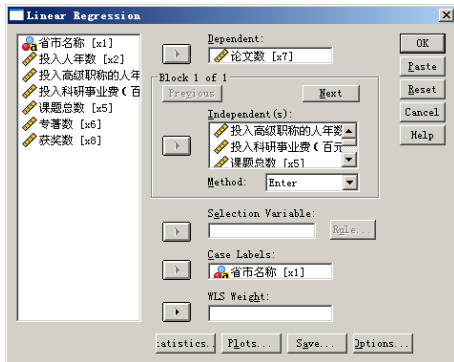


FIG: 线性回归对话框



## 回归分析的简单基本操作及参数的估计 II

- ② 选择被解释变量和解释变量，“Dependent”为因变量，亦即被解释变量；“independent”为自变量，亦即解释变量，如果需要对不同的变量采用不同的引入方法(逐步回归方法)，例如对前两个变量用强迫引入法，可以先选前两个变量进入，选 method 为 enter, 组成一个 block, 点击 next, 选其他的变量组成下一个 block, 对每个 block 可以定义不同的引入方法。
- ③ 选择个案标签，被选中的变量可以用来在做图时标记不同的点。
- ④ 选择加权二乘法变量，选择权变量 WLS weight 变量，进行加权二乘法时需要，一般为观测重数。
- ⑤ 执行操作，点击 OK。



# METHOD选项介绍

## METHOD选项介绍

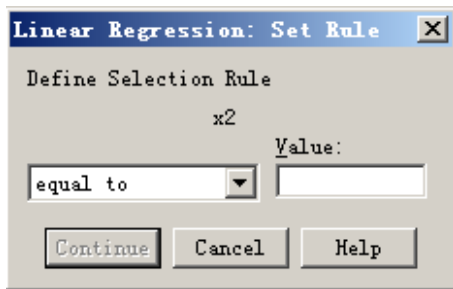
- Enter 强迫进入法，默认选项，全部变量一次性进入回归模型
- Stepwise 逐步法，在每次引入变量时，概率F值最小的变量进入回归方程，如果引入方程的变量F值大于给定值，则剔除，如果没有变量被引入或剔除时，终止回归过程
- Remove 剔除变量，将所有不进入方程模型的被选变量一次性剔除
- Backward 向后法，选取所有的变量进入模型，逐步将p值较大的变量剔除，找到最优模型
- forward 向前法，被选变量依次进入模型，首先引入与因变量相关性最大且符合标准的变量(p值小于给定值)，当无变量被引入时，过程终止。



# RULE选项

## RULE 选项

将列表中变量选入"Selection Variable"列表框中, 该变量用于指定分析个 案的选择规则。单击"Rule"按钮, 打开设定规则的对话框, 在define Selection Rule 中左侧的下拉菜单中给出设定临界规则的选项





# 回归方程的统计检验

通过样本数据建立回归方程之后一般不能立即用于对实际问题的分析和预测，通常要进行各种统计检验，包括回归方程的拟合优度检验、回归方程的显著性检验、回归系数的显著性检验、残差分析等。



## 回归方程的拟合优度检验 I

由于回归方程反映的是解释变量 $x$ 的不同取值变化对被解释变量 $y$ 的线性影响规律，由 $x$ 的变化引起的 $y$ 的变差的平方和 $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 称为回归平方和(SSA)，而由随机因素导致的 $y$ 的变差平方和 $\sum_{i=1}^n (\hat{y}_i - y_i)^2$ 通常称为剩余平方和(SSE)，也称为残差平方和。且有总平方和SST的关系式

$$SST = SSA + SSE$$

讨论:

- ① 如果所有的点都在直线上，那么此时剩余平方和为0，总平方和SST由回归平方和构成，此时解释变量的值可以完全确定被解释变量
- ② 如果所有的点都是散乱的，总平方和中回归平方和占的比例很小，那么说明解释变量的作用很小
- ③ 由此可知，回归方程能够解释的变差在总变差中的比例越大，说明回归方程拟合得越好



## 回归方程的拟合优度检验 II

由此可以得出统计量

- 对于一元线性回归模型

$$R^2 = \frac{SSA}{SST} = 1 - \frac{SSE}{SST}$$

即为回归平方和在总平方和中的比例， $R^2$ 的取值在0到1之间， $R^2$ 越接近于1，说明回归方程的拟合程度越好，反之 $R^2$ 越接近于0拟合程度越差

- 对于多元回归模型，采用 $\bar{R}^2$ 统计量，称为调整的判定系数或调整的决定系数，定义为

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$



# 线性回归方程的显著性检验 I

线性回归方程显著性假设为

$$H_0 : \beta = 0$$

即检验回归参数是否为零，如果为0，则说明被解释变量和解释变量之间不具有线性关系，回归方程没有意义，线性回归方程不能够解释被解释变量和解释变量之间的关系。回归方程显著性检验的基本出发点与拟合优度检验非常相似。采用平均的SSA和平均的SSE的比值作为统计量。

- 对于一元线性回归方程，该检验方法的统计量为

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - 2)}$$

F统计量服从参数为(1,  $n - 2$ )的F分布



## 线性回归方程的显著性检验 II

- 对于多元回归方程，该检验方法的统计量为

$$F = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n - p - 1)}$$

其中 $p$ 为解释变量的个数。F统计量服从自由度参数为 $(p, n - p - 1)$ 的F分布。

SPSS将自动计算F统计量的值和 $p$ 值，如果 $p$ 值很小，小于给定的显著性水平，则拒绝原假设，认为被解释变量 $y$ 与解释变量 $x$ 之间的线性关系显著，可以用线性模型描述和反映他们之间的关系；反之，接受原假设，认为线性关系不显著。但此时不能认为变量之间独立，应该通过散点图找出解释变量和被解释变量之间的非线性关系



## 回归系数的显著性检验 I

回归系数的显著性检验的主要目的是研究回归方程中的每个解释变量与被解释变量之间是否存在显著的线性关系，也就是研究解释变量能否有效地解释被解释变量的线性变化。回归系数的显著性检验的原假设为

$$\beta_i = 0, \text{对固定的} i$$

- 在一元线性回归中检验统计量为t统计量,其构造如下

$$t = \frac{\beta_i}{\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

其中,  $\hat{\sigma}$  为回归方程的标准误差, 满足

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



## 回归系数的显著性检验 II

是误差平方和SSE的平均，反映了回归方程无法解释的y的变动部分。零假设条件下，t统计量服从参数为(n-2)的t分布。

- 在多元线性回归中检验统计量的构造如下

$$t_i = \frac{\beta_i}{\frac{\hat{\sigma}}{\sqrt{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2}}}, i = 1, \dots, p$$

其中 $t_i$ 服从参数为n-p-1的t分布，

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

SPSS将自动运算相应的t值和p值。我们可以根据p值进行推断



# 关于显著性检验的注记

回归方程显著性检验与回归系数的显著性检验的作用不同

- 回归方程显著性检验只检验所有偏回归系数是否同时为0
- 如果不同时为0，并不排除存在某些系数为0 的情况
- 所以，回归方程显著后需要对每个偏回归系数作检验，剔除不显著的解释变量，





# 残差分析

所谓残差是指由回归方程计算所得的预测值(估计值)和实际观察值之间的差距, 定义为

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip})$$

是回归分析中 $\varepsilon_i$ 的估计值, 多个 $e_i$ 形成的序列称为残差序列。残差分析是回归方程检验中的重要组成部分, 其出发点是如果回归方程能够较好的反映被解释变量的特征和变化规律, 那么残差序列中应不包含明显的规律性和趋势性。

## 残差分析的主要任务

- 检验残差是否为均值为0的正态分布
- 残差是否是相互独立的
- 残差的方差是否相同(异方差检验)
- 探测样本中的异常值和强影响点



## 残差均值为0 的正态性分析

可以通过绘制残差图进行分析，如下图，可以看出第一个图形中的残差近似均值为零，第二个图形的残差是由偏移的

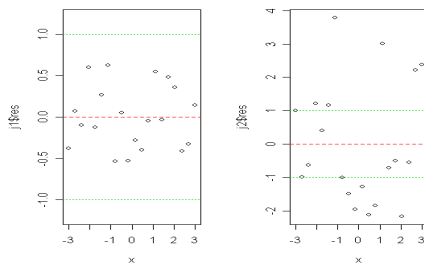


FIG: 残差分析图形一，正态性分析



# 残差的独立性分析 I

残差序列的独立性也是回归模型所要求的，残差序列应满足 $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$ ，残差独立性分析可以通过三种方式实现：

- 其一是绘制残差序列的图形,此时可以看出残差序列存在一定的相关性
- 计算残差的相关系数,其计算公式为

$$\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sqrt{\sum_{t=2}^n e_t^2 \sum_{t=2}^n e_{t-1}^2}}$$

自相关系数的取值范围在-1到1之间，接近于1 说明序列存在正相关，接近于-1 说明序列存在负相关。



## 残差的独立性分析 II

- DW(Durbin-Watson)检验

其零假设为总体的自相关系数与0没有显著差异。采用的统计量为

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=2}^n e_t^2}$$

DW的取值在(0,4)之间,  $DW = 4$ 时, 序列存在完全负相关,  $DW$ 等于2时, 序列不相关,  $DW$ 等于0时, 序列存在完全正相关

如果残差序列存在自相关, 说明回归方程不能充分说明被解释变量的变化规律, 需要重新考虑。



# 探测样本中的异常值和强影响点

对被解释变量 $Y$ 中异常值的探测方法一般有如下几种

- 标准化残差
- 学生化残差
- 剔出某个观测后的残差
- 杠杆值
- 库克距离
- 标准化回归系数的变化和标准化预测值的变化



## 曲线回归分析和曲线拟合

变量间的相关关系的分析中，变量之间的关系并不总是表现为线性关系，非线性关系也是常见的，通过绘制散点图可以粗略考察这种非线性关系。变量之间的非线性关系可以划分为本质线性和本质非线性关系

- 所谓本质线性关系是指变量关系形式上虽非线性关系，但可以通过变量变换转化为线性关系，并通过线性回归分析建立线性模型。如：

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

取 $x_1 = x^2$ ，上述模型可以转化为

$$y = \beta_0 + \beta_1 x + \beta_2 x_1$$

- 本质非线性关系是指变量关系不仅形式上呈非线性关系，而且也无法通过变量变换转化为线性关系，最终无法通过线性回归分析建立线性模型。
- 本节的曲线估计是解决本质线性关系问题的



# 常见的本质线性模型 I

- 二次曲线(Quadratic)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \text{ 同上可令 } x_1 = x^2$$

- 复合曲线(Compound)

$$y = \beta_0 \beta_1^x, \text{ 两边取对数得 } \ln(y) = \ln(\beta_0) + \ln(\beta_1)x$$

- 增长曲线(Growth)

$$y = e^{\beta_0 + \beta_1 x}, \text{ 变换为 } \ln(y) = \beta_0 + \beta_1 x$$

- 对数曲线(Logarithmic)

$$y = \beta_0 + \beta_1 \ln(x) \text{ 变换为 } y = \beta_0 + \beta_1 x_1 (x_1 = \ln(x))$$

- 三次曲线(Cubic)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 (x_1 = x^2, x_2 = x^3)$$



## 常见的本质线性模型 II

- S曲线

$$y = e^{\beta_0 + \beta_1/x} \left( \text{令 } x_1 = \frac{1}{x} \right)$$

- 指数曲线(Exponential)

$$y = \beta_0 e^{\beta_1 x}, \text{ 转化为 } \ln(y) = \ln(\beta_0) + \beta_1 x$$

- 逆函数(Inverse)

$$y = \beta_0 + \beta_1/x$$

- 幂函数(Power)

$$y = \beta_0 x^{\beta_1}$$

- 逻辑函数(Logistic)

$$y = \frac{1}{1/\mu + \beta_0 \beta_1^x}$$

变换可得

$$\ln\left(\frac{1}{y} - \frac{1}{\mu}\right) = \ln(\beta_0) + \ln(\beta_1)x$$





# SPSS曲线估计

在SPSS曲线估计中，首先在不能明确究竟哪种模型更接近样本数据时，可以在上述多种可选模型中选择几种，然后SPSS自动完成模型的参数估计，并输出回归方程显著性检验的F值和概率p值，判定系数 $R^2$ 等，最后以判定系数为主要依据选择其中的最优模型，并进行预测分析等



# 曲线估计的基本操作



## 二项LOGISTIC回归

利用多元回归方法分析变量之间的关系或进行预测时的一个基本要求是，被解释变量应为连续行定距型变量(Scale)。然而 在实际应用中未必都能够得到较好的满足。

### EXAMPLE (购买小轿车)

在对小轿车消费群体特点的分析和预测研究中，可以根据历史数据，建立关于购买小轿车的多元回归模型，可能将职业、年收入、年龄、性别等变量纳入模型，并希望通过模型预测具有某种特定特征的客户是否会购买小轿车。这个多元回归模型的被解释变量 设为购买与否(购买记为1，不购买记为0)，是个纯粹的二值变量，显然不满足多元回归分析的要求，而在数据分析，特别是社会 科学研究中，这种情况经常出现。



# 解释变量去二值时一般多元回归模型的问题

## 误差出现的问题

- ❶ 残差不再满足  $E(\varepsilon) = 0, \text{Var}(\varepsilon) = \sigma^2$  的假设条件
- ❷ 残差不再服从正态分布
- ❸ 被解释变量的取值空间受限制

由此可见，当被解释变量是0/1二值品质变量时，无法直接采用一般的多元线性回归模型建模，通常采用Logistic回归。 **Logistic回归是多元回归方法不断发展的成果**



# LOGISTIC模型的思想 I

- ① 用解释变量 $X$ 的取值 $x_i$ 解释被解释变量 $Y$ 取1的概率,  $P(Y = 1) = \beta_0 + \beta_i x_i$
- ② 设 $P = P(Y = 1)$ , 则 $P$ 的取值范围在 $(0, 1)$ , 不能够进行回归分析, 做变换

$$\Omega = \frac{P}{1 - P}$$

此时 $\Omega$ 的变化与 $P$ 同方向, 但是只取 非负值

- ③ 做变换

$$\ln(\Omega) = \ln\left(\frac{P}{1 - P}\right)$$

$\ln(\Omega)$ 称为Logit  $P$ , 经过这样的转换后 $\text{Logit} P$ 与 $P$ 的变化同方向, 且取值于 $(-\infty, \infty)$ , 与一般的 线性回归模型中的被解释变量的取值范围吻合。



## LOGISTIC模型的思想 II

经过上述的转换后，可以得到模型

$$\text{Logit}P = \beta_0 + \beta_i x_i$$

计算可得：

$$P = \frac{1}{1 + \exp -(\beta_0 + \beta_i x_i)}$$

上式正是Logistic函数，是典型的增长函数，正好体现了概率P和被揭示变量之间的关系。



## 二项LOGISTIC回归方程中回归系数的含义

当Logistic回归模型的回归系数确定后，将其代入到 $\Omega$ 的函数中，有

$$\Omega = \exp(\beta_0 + \beta_i x_i)$$

当其他解释变量保持不变而研究 $x_1$ 变化一个单位对 $\Omega$ 的影响，可将新的发生比设为 $\star\Omega$ ，则：

$$\star\Omega = \exp(\beta_1 + \beta_0 + \beta_i x_i) = \Omega \exp(\beta_1)$$

由此可知，当其他解释变量保持不变时， $x_i$ 每增加一个单位将引起发生比扩大 $\exp(\beta_i)$ 倍，当回归系数为负时，发生比缩小。



## 二项LOGISTIC回归方程的检验

Logistic回归方程的参数求解采用极大似然估计法。基于总体的分布密度函数和样本信息的基础上，求解在似然函数值最大下的未知参数的估计值。在该原则下得到的模型，其产生的样本数据的分布与总体分布相近的可能性最大。

因此似然函数的函数值实际上也是一种概率值，反映了在所确定的拟合模型为真时该模型能够较好地拟合样本数据的可能性。所以似然函数的取值在0~1之间。

### 回归方程的显著性检验

Logistic回归方程显著性检验的目的是检验解释变量全体与Logit  $P$ 的线性关系是否显著，是否可以用线性模型拟合。其零假设是回归系数同时为0，解释变量全体与Logit  $P$ 的关系不显著。





## 显著性检验的基本思想

如果方程中的诸多解释变量对Logit P 的线性解释有显著意义，那么必然会使回归方程对样本数据的拟合得到显著提高，可采用对数似然函数的比值来测度拟合程度是否有提高。设系数全部为0时的似然函数的值为 $L_0$ ，解释变量引入回归方程后对数似然函数值为 $L$ ，取对数似然比

$$LR = \frac{L}{L_0}$$

如果LR的值与1没有什么差别，说明模型不显著，如果远大于1，说明当前模型中的解释变量的全体对Logit P有显著贡献。实际中采用的统计量为

$$-2 \ln\left(\frac{L_0}{L}\right)$$

渐进服从卡方分布 $Q\chi^2$



## 回归系数的显著性检验

Logistic 回归系数显著性检验的目的是逐个检验模型中各解释变量是否与Logit P有显著的线性关系。 Logistic 回归系数的显著性检验采用的统计量是Wald 检验统计量，定义为

$$Wald_i = \left( \frac{\hat{\beta}_i}{S_{\beta_i}} \right)^2$$

其中， $\hat{\beta}_i$ 为回归系数， $S_{\beta_i}$ 为回归系数的标准误差。Wald检验统计量在零假设条件下服从 $\chi^2(1)$ 。零假设 $H_0$ 为 $\beta_i = 0$  SPSS将自动计算所有的Wald统计量的值和相应的p值。



# 回归方程的拟合优度检验

在Logistic回归分析中，拟合优度可以从两方面来考察

- 回归方程能够解释被解释变量变差的程度，如果方程可以解释被解释变量的较大部分变差，则说明拟合优度高，反之说明拟合优度低
- 由回归方程计算出的预测值与实际值之间的吻合程度，即方程的错判概率是低还是高。

常见的指标为

- ① Cox Snell  $R^2$ 统计量，类似与线性回归中的总相关系数
- ② Nagelkerke  $R^2$ 统计量，为修正的 Cox Snell  $R^2$ 统计量
- ③ 错判矩阵
- ④ Hosmer-Leemeshow统计量 解释变量为定矩型变量时多采用



## 残差分析

Logistic回归中可以利用以下残差指标进行残差分析，主要包括非标准化残差、标准化残差、Logit残差等，

- 非标准化残差定义为

$$e_i = y_i - P_i(y = 1|x_i)$$

- 标准化残差定义为

$$\text{Standard } e_i = \frac{y_i - n_i P_i}{\sqrt{n_i P_i (1 - P_i)}}$$

其中  $n_i$  表示解释变量取特定值的样本个数

- Logit 残差定义为

$$\text{Logit } e_i = \frac{e_i}{P_i(1 - P_i)}$$




## 二项LOGISTIC回归中的虚拟变量

- 被解释变量的变化会受到非定矩型的品质变量的影响
  - 客户是否购买小轿车不仅受年收入和年龄的影响，还可能受到诸如性别、职业等影响
- 。 品质型数据通常不能像定矩型变量那样直接作为解释变量，因此一般需要将其转化成虚拟变量或哑变量(Dummy Variable)



## 二项LOGISTIC回归的基本操作

- 1 选择菜单Analysis  $\Rightarrow$  Regression  $\Rightarrow$  Binary Logistic
- 2 选择一个被解释变量选择到Dependent框，把一个或多个解释变量选择到Covariates框。
- 3 method中选择解释变量的筛选策略
- 4 单击 Select 按钮，选择一个变量作为条件到Selection Variable框中，并单击Rule按钮给定一个判断条件。
- 5 如果解释变量为非定矩型数据，可单击Categorical 按钮，选择属性变量，定义哑变量 



# 定义哑变量

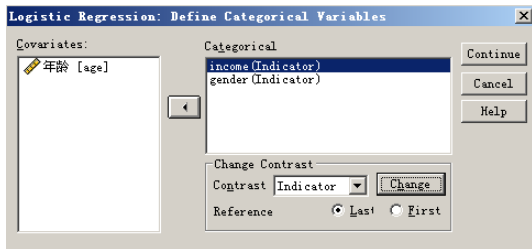


FIG: 哑变量的定义

change contrast 选择框中选择不同的对照方法可以定义哑变量，  
具体参见Help



## 二项LOGISTIC回归的其他操作

### OPTION选项

- Statistics and Plots 框中:
- Display框: 可选择输出每一步结果或只输出最终结果
- Probability for Stepwise 框: 指定进入或提出变量的显著性水平
- Classification Cutoff: 设置概率分界值
- Maximum iterations: 指定最多的循环次数





## 其中的STATISTICAL AND PLOT 选项

- Classification Plots表示绘制被解释变量实际值与预测分类值得关系图
- Hosmer-Lemesho goodness-of-fit表示输出Hosmer-Lemeshow拟合优度指标
- Casewise Listing of Residuals表示输出各样本数据的非标准化残差、标准化残差等指标
- CI for exp(D) 表示输出风险比默认为95%的置信区间



# SAVE 选项

- Predicted Value框中，Probabilities表示保存被解释变量取1的预测概率值，Group membership 表示保存分类预测值
- Residuals 和 Influence 表示保存残差、库科距离、杠杆值。



## 二项LOGISTIC回归的应用举例

参看数据



# 非参数检验简介

非参数检验方法主要涉及以下方面

- 单样本非参数检验
- 两独立样本非参数检验
- 两配对样本非参数检验
- 多独立样本非参数检验
- 多配对样本非参数检验

这里我们主要介绍单样本非参数检验和两独立样本非参数检验。



# 单样本非参数检验

## EXAMPLE

拿到一批样本数据后，往往希望了解样本来自的总体是否与某个已知的分布相吻合。可以通过做直方图，P-P图，Q-Q图的方法作粗略的判断，也可以通过非参数检验的方法来实现。

另外非参数检验还可以用来检验变量的随机性，比如01序列中0和1是否是随机的出现的。



# 总体分布的卡方检验—基本思想

## 卡方检验的理论依据是:

如果从一个随机变量 $X$ 中随机抽取若干个观察样本, 这些观察样本落在 $X$ 的 $k$ 个互不相交的子集中的观察频数 应该服从一个多项分布。基于这一思想, 对变量 $X$ 的总体分布的检验就可以对各个观察频数的分析入手。



# 卡方检验统计量的构造

在零假设成立的条件下(即样本服从已知的理论分布如正态, 指数等), 通过理论分布可以计算变量值落在第 $i$ 个子集中的概率为 $p_i, i = 1, \dots, k$ , 于是由样本得出的相应的期望频数就应该是 $np_i$ 。此期望频数代表了零假设成立时的理论分布。此时从样本中得出的观测频数记为 $f_i$ , 构造如下统计量

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$$

在零假设成立时, 有统计学理论可以得出 $\chi^2 \sim \chi^2(k-1)$ , 自由度为 $n-1$ 的卡方分布, 如果样本确实服从此理论分布, 则卡方统计量的值偏小。SPSS将自动计算统计量的值和尾概率。



# 卡方检验的基本操作





## 二项分布检验的基本思想

在现实生活中有很多数据的取值是二值的，如产品合格不合格等，通常将这样的二值分别用0和1表示。如果随机变量 $X$ 取1的概率为 $p$ ，则取0的概率为 $1 - p$ ，形成二项分布。SPSS的二项分布检验正是要通过样本数据检验样本来自的总体是否服从概率为 $p$ (指定值)的二项分布。



## 二项分布检验的统计量构造

- 在小样本条件下，采用精确检验方法，计算n次试验中某类出现的次数小于等于x的概率

$$P(X \leq x) = \sum_{i=0}^x C_n^i p^i q^{n-i}$$

- 大样本条件下，采用Z检验统计量，在零假设成立的条件下Z统计量近似服从正态分布

$$Z = \frac{x \pm 0.5 - np}{\sqrt{np(1-p)}}$$

上式进行了连续性校正，当 $x < n/2$ 时加0.5，当 $x > n/2$ 时减0.5。SPSS将自动计算上述精确概率值和近似概率值



# 单样本KOLGOROV-SMIRNOV检验的基本思想

K-s检验是以俄罗斯数学家Kologorov和Smirnov命名的一种非参数检验方法，该方法的思想是当样本数比较多时，经验分布函数接近于理论分布函数，适用于探索连续型随机变量的分布。

**K S检验的零假设为：样本来自的总体与指定的理论分布无显著差异**SPSS的理论分布主要包括正态分布，均匀分布，指数分布和泊松分布等。



# K-S检验统计量的构造

首先定义经验分布函数，表示样本中小于等于 $x$ 的观测频率

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq x)}$$

其中 $I_{(X_i \leq x)}$ 为示性函数，取值为1或0。由概率统计的知识可以得到，当样本数比较多时

$F_n(x)$ 以概率收敛到 $F(x)$

于是可以选取统计量

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|$$

当原假设成立时，上述统计量服从既定的分布。在SPSS中，自动计算 $\sqrt{n}D_n$ 的值和相应的尾概率值



# K-S检验



# 变量随机性检验的基本思想

变量值随机性检验通过对样本变量的分析，实现对总体变量值出现是否随机进行检验。

例如在投硬币时，如果1表示出现正面，0表示出现反面，在进行了若干次投币后，将会得到一个以1，0组成的变量值序列，这时可能会分析“硬币出现正反面是否随机”这样的问题？游程检验正是解决这种问题的方法。



# 游程检验

## 游程的含义:

假定下面是由0和1组成的一个这种变量的样本:

0 0 0 0 1 1 1 1 1 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0

其中相同的0（或相同的1）在一起称为一个游程（单独的0或1也算）。这个数据中有4个0组成的游程和3个1组成的游程。一共是 $R=7$ 个游程。其中0的个数为 $m=15$ ，而1的个数为 $n=10$ 。

## 游程检验的原理:

判断数据序列是否是真随机序列。该检验的原假设为数据是真随机序列，备择假设为非随机序列，在原假设成立的情况下，游程的总数不应太多也不应太少。



## 利用游程数构造统计量

游程是一个具有独特抽样分布的统计量。如果设 $n_1$ 为出现1的个数， $n_2$ 为0出现的个数。当 $n_1, n_2$ 较大时，游程抽样分布的均值为 $\mu_r = \frac{2n_1n_2}{n_1+n_2}$ ，方差为 $\sigma_r^2 = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$ ，在大样本时，游程近似服从正态分布，即

$$Z = \frac{r - \mu_r}{\sigma_r}$$

其中 $r$ 为游程数。SPSS将自动计算 $Z$ 统计量的值和概率 $p$ 值得近似值。SPSS操作时可以对非二值变量选择分割值。





# 两独立样本的非参数检验简介

两独立样本的非参数检验正是在总体分布不甚了解的情况下，推断两个总体的分布是否存在显著差异的方法。

SPSS中提供了多种两样本的非参数检验方法

- 曼-惠特尼检验(U检验)
- K-S检验
- W-W游程检验
- 极端反应检验



# 两独立样本的MANN-WHITNEY检验的基本思想

曼-惠特尼检验通过比较两样本的秩进行，其基本步骤是

- ① 将两组样本数据 $(X_1, X_2, \dots, X_m), (Y_1, Y_2, \dots, Y_n)$ 混合并按升序排序，得到每个数据各自的秩
- ② 分别求出X和Y的秩的平均值 $W_x/m, W_y/n$ ,对两个平均秩进行比较。
- ③ 显然，如果两个平均秩差别较大，则应是一组样本秩偏小，另一组样本秩偏大的结果，此时说明一组样本的值偏小，零假设 不成立



## 曼-惠特尼统计量的构造

取

$$W = \sum_{i,j} I_{(X_i \leq Y_j)}$$

其中， $I_{(X_i \leq Y_j)}$ 为示性函数，取值为0或1。可以看出如果所有的X比Y小，W取大值，反之取小值。



# 曼-惠特尼统计量的分布

曼-惠特尼统计量取为

$$U = W - \frac{1}{2}k(k+1)$$

k为对应样本组的样本数。小样本情况下，U统计量服从曼-惠特尼分布。大样本下，U统计量近似服从正态分布，取

$$Z = \frac{U - \frac{1}{2}mn}{\sqrt{\frac{1}{12}mn(m+n+1)}}$$

SPSS将自动计算统计量的值和p值。



## 两独立样本K-S检验基本思想

与单样本K-S检验的区别主要是:

- 首先, 将两组样本混合并按升序排序
- 分别计算两组样本秩的累积频数和累积频率
- 最后计算两组累积频率的差, 得到差值序列并得到 $D$ 统计量

SPSS将自动计算在大样本情况下 $\sqrt{n}D$ 的观测值和概率 $p$ 值。关于两独立样本的游程检验和极端反应检验 不再赘述。



## 两独立样本的游程检验的基本思想

- 两独立变量的游程检验用来检验两独立样本来自的两总体的分布是否存在显著差异，其零假设 $H_0$ 为：两独立样本来自的两总体的分布无显著差异。
- 两独立样本的游程计算依赖于秩。

### 两样本游程的计算

- 首先将两组样本混合按升序排序
- 然后按照前面讨论的计算游程的方法计算
- 易见：如果两组样存在较大差距，那么游程数会相对较少，如果游程数比较大，则应是两组样本值混合充分的结果
- 根据游程数计算统计 $z$ ，该统计量近似服从正态分布
- 缺点：只考虑大小排序，对取值上量的差别关注不大，没有充分利用样本信息



# 极端反应检验的基本思想

极端反应检验的从另一个角度检验两独立样本所来自的两总体的分布是否存在显著差异。其零假设 $H_0$ 为：两独立样本来自的两个总体的分布无显著差异。

## 基本思想：

将一组样本作为控制样本，另一组样本为实验样本，以控制样本为参照，检验实验样本相对于控制样本是否出现了极端反应。如果实验样本没有出现极端反应，则认为两总体分布存在显著差异。具体过程为：

- 将两组样本按升序排序
- 求出控制样本的最小秩 $Q_{min}$ 和最大秩 $Q_{max}$ ，并计算出跨度SPAN

$$S = Q_{max} - Q_{min} + 1$$

极端反应检验注重对跨度的分析，如果跨度小，则是两样本数据无法充分融合，一组样本值显著大于另一组，如果跨度较大，则应是两组样本数据充分混合



# 两独立样本非参数检验的基本操作





## 两配对样本的McNEMAR检验 的基本思想

McNemar 检验是一种变化显著性检验，将研究对象在实验前后的变化是否显著。其零假设为：两配对样本来自的总体分布无显著差别。



# 聚类分析简介

- 物以类聚、人以群分;
- 但根据什么分类呢?
- 如要想把中国的县分类,就有多种方法 可以按照自然条件来分,比如考虑降水、土地、日照、湿度等, 也可考虑收入、教育水准、医疗条件、基础设施等指标;
- 既可以用某一项来分类,也可以同时考虑多项指标来分类。



# 变量聚类 and 个案聚类

- 对一个数据，既可以对变量(指标)进行分类(相当于对数据中的列分类)，也可以对观测值(事件，样品)来分类(相当于对数据中的行分类)。
- 当然，不一定事先假定有多少类，完全可以按照数据本身的规律来分类。
- 本章要介绍的分类的方法称为聚类分析 (cluster analysis)。对变量的聚类称为R型聚类，而对观测值聚类称为Q型聚类。它们在数学上是无区别的



# 聚类分析中亲疏程度的度量方法

个体之间的亲疏程度的测度一般有两个角度

- ① 个体间的相似程度，用简单相关系数或等级相关系数刻画
- ② 个体间的差异程度，用某种距离来测度

个体或变量间的亲疏程度的度量可以通过SPSS中的距离分析来实现。



# 距离分析的基本概念

距离分析(Distances)是对观测变量之间相似或不相似程度的一种测度，是计算一对变量之间或一对观测变量之间的广义的距离。这些相似性或距离测度可以用于其他分析过程，如因子分析，聚类分析等。

在距离分析过程中，主要利用变量间的相似性测度(Similarities)和不相似性测度(Dissimilarities)度量两者之间的关系。



# 相似性测度

两变量之间的相似性，可以用相关系数等度量，

- 针对定距型数据有pearson相关系数和夹角余弦距离等
- 对二值变量的相似性测度主要包括简单匹配系数(Simple matching), Jaccard相似性指数, Hamann相似性测度等20余种
- 另外，相似性测度可以用于因子分析、聚类分析等模块



# 不相似性测度

- 对定距型变量间距离描述的统计量，主要有欧式距离，平方欧式距离，契比雪夫(Chebychev)距离、Block距离、Minkowski距离
- 对于定序型变量之间距离的描述，主要有平方不相似测度(Chi-Square measure)和Phi-Square不相似测度
- 对于二值变量之间的距离描述，主要有欧式距离，平方欧式距离，Lane and Williams不相似测度等



# 聚类分析的几点说明

- ① 所选变量应符合聚类的有求  
例如，如果希望按照各学校的科研情况进行分类，应该把科研相关变量选入分类变量，而学生人数，校园面积等变量应该剔除。
- ② 各变量的变量值不应有数量级的差异，如投入经费以万元计还是用元计会对分类产生影响，为消除数量级对分类的影响，可以事先对变量进行标准化变换。
- ③ 各变量之间不应有较强的线性相关关系

常见的聚类分析方法有层次聚类和K-Means快速聚类方法，以下分别介绍之。。





# 层次聚类方法

## 层次聚类有两种类型

### ① Q聚类

是对样本进行聚类，使具有相似特征的样本聚集在一起

### ② R聚类

是对变量进行聚类，使具有相似特征的变量聚集到一起

## 层次聚类的方式也分为两种

### ① 凝聚方式分类

### ② 分解方式聚类

SPSS中的层次聚类采用的是凝聚方式。



## 层次聚类的凝聚方式分类

所谓凝聚方式指的是：先将每一个样本或变量都看作为一个单独的类，比较两个类之间的距离，最近的两个样本或变量**凝聚**为一个类。然后计算两个类间的距离，依次类推下去，最后凝聚为一个类，而操作者可以根据自己的意愿选择究竟选择多少个类最合适。反过来，分解聚类法和凝聚方法正好相反，不再赘述。



谢谢!

欢迎提问

