# 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

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* SPSS简介
* SPSS产品
* SPSS特点
* SPSS模块
* SPSS基础操作
* SPSS的文件菜单
* SPSS数据文件的特点
* SPSS数据录入格式要求
* SPSS的变量名
* SPSS变量类型
* SPSS变量尺度
* SPSS的值和标签
* SPSS的缺失值
* 导入文本文件
* 导入excel文件
* 手动录入数据
* 变量赋值
* 重新编码为（相同）不同变量
* 数值变量的离散化分段
* 计算个案内值的出现次数
* 记录排序
* 记录拆分
* 记录筛选
* 记录加权
* 分类汇总
* 多个数据文件的合并
* 标识重复个案
* 数据验证

### 【重点】

* SPSS变量尺度
* SPSS的缺失值
* 导入文本文件
* 导入excel文件
* 手动录入数据
* 重新编码为（相同）不同变量
* 数值变量的离散化分段
* 计算个案内值的出现次数
* 记录排序
* 记录拆分
* 记录筛选
* 记录加权
* 分类汇总
* 多个数据文件的合并
* 标识重复个案

### 【难点】

* 重新编码为（相同）不同变量
* 数值变量的离散化分段
* 计算个案内值的出现次数
* 记录排序
* 记录拆分
* 记录筛选
* 记录加权
* 分类汇总
* 多个数据文件的合并
* 标识重复个案

## 复习巩固作业讲解

无

## 本次课程任务讲解

### 【知识点1】SPSS简介

SPSS软件是世界上应用最广泛的专业统计软件之一，在全球约有25万用户，分布于通信、医疗、银行、证券保险、制造、商业、市场研究和科研教育等众多领域和行业，全球500强中约有80%的公司使用SPSS，而在市场研究和市场调查领域则拥有超过80%的市场占有率，和SAS被称为当今最权威的两大统计软件。

SPSS实际上是该软件的简称，其全称则发生过几次变化，最早为Statistical Package for Social Sciences，意为“社会科学统计软件包”；后来随着SPSS产品服务领域扩大，SPSS公司于2002年更改为“统计产品与服务解决方案”（ Statistical Product and Service Solutions）;2009年，SPSS被IBM收购，重新将软件命名为IBM SPSS Statistics 。

* 11版：重新设计了软件界面，加入了混合线性模型等新方法。
* 12版：图形输出更改为目前使用的系统，也是首个提供简体中文版的版本。
* 13版：真正的经典之作，还有很多用户在使用。
* 14版：首次可以同时打开多个数据文件，提供了现在已成为主要绘图界面的“图表生成器”界面。
* 15版：提供可直接调用Python等语言编写代码。对硬件要求提高了。
* 16版：用Java重写了整个用户界面，操作更加灵活。
* 17版：首次提供了包括简体中文在内的多语言界面。
* 18版：更新了一系列统计方法，数据管理、用户界面等，每个模块可以独立运行。
* 19版：IBM并购后的第一个版本，并未做较大更新。
* 20版：该版本修正了上一版本中的问题和改进。

### 【知识点2】SPSS产品

目前SPSS产品已形成4大产品系列构成了完整的产品线，SPSS的定位就是针对常规统计分析与简单的数据挖掘。

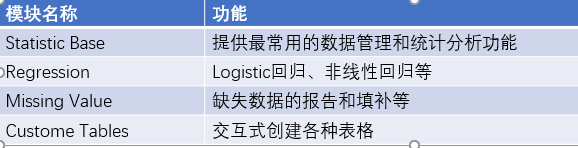
* Data Collection Family ： 定位中低端的数据采集与报告需求，是一个完整的技术平台，支持从创建调查到收集数据，再到报告的整个调查研究的生命周期。
* Statistics Family ： 定位中端的统计分析服务需求，有原先的SPSS软件构成，但Base模块不再是必须，原先的每个附加模块现在都可以独立安装和运行。
* Modeling Family ： 有原先的Clementine发展而来，现更名为IBM SPSS Modeler 主攻高端的数据挖掘与商务智能需求领域。
* Deployment Family ： 相对是前3个产品系列的整合与后台支持，可以把市场调研、统计分析技术、数据挖掘技术以及报表技术整合到一个平台中。

### 【知识点3】SPSS特点

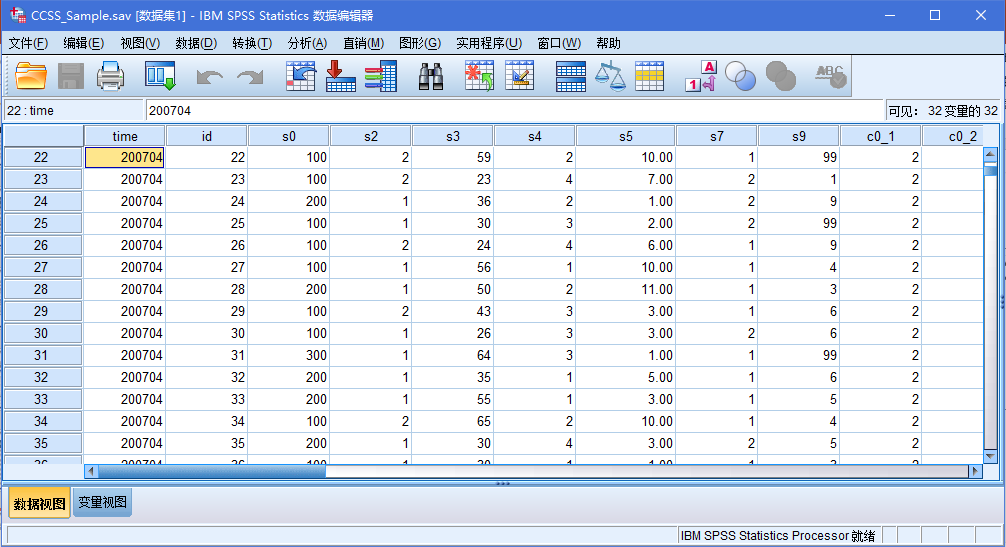
* **功能强大**：SPSS囊括了各种成熟的**统计方法和模型**，为统计分析用户提供了全方位的统计学算法。
* **兼容性好**：在数据方面，不仅可以在SPSS中直接进行数据录入工作，还可以将日常工作中常用到的Excel表格数据、文本格式数据导入SPSS中进行分析，从而节省了相当大的工作量。
* **易用性强**：SPSS之所以有广大的用户群，不仅因为它是一种权威的统计学工具，提供了强大的统计功能，也因为它是一种非常简单易用的软件。
* **扩展性高**：SPSS直接和R语言进行对接，通过直接调用R语言的各种统计模块，直接实现了对最新统计方法的调用。

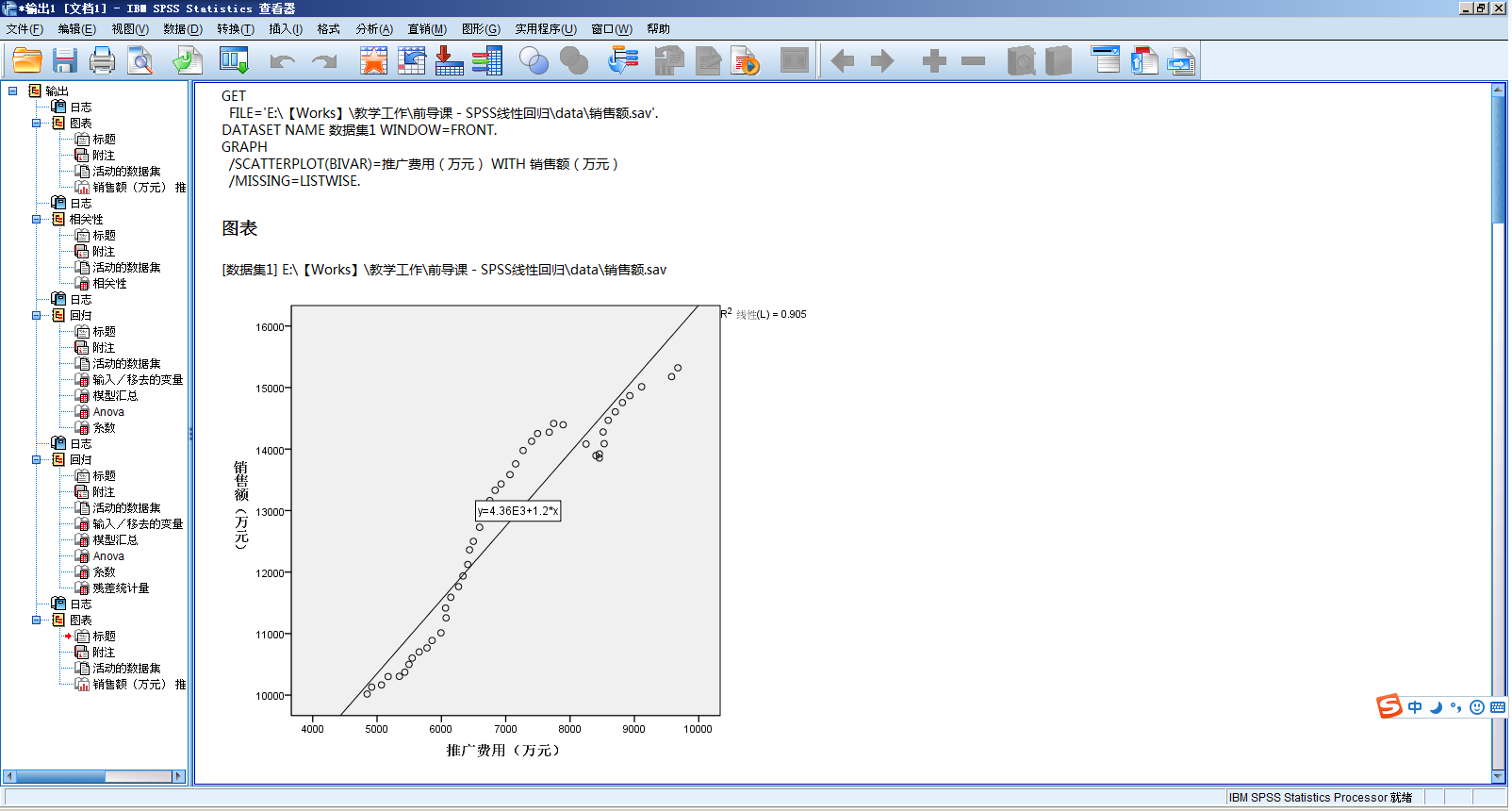
### 【知识点4】SPSS模块

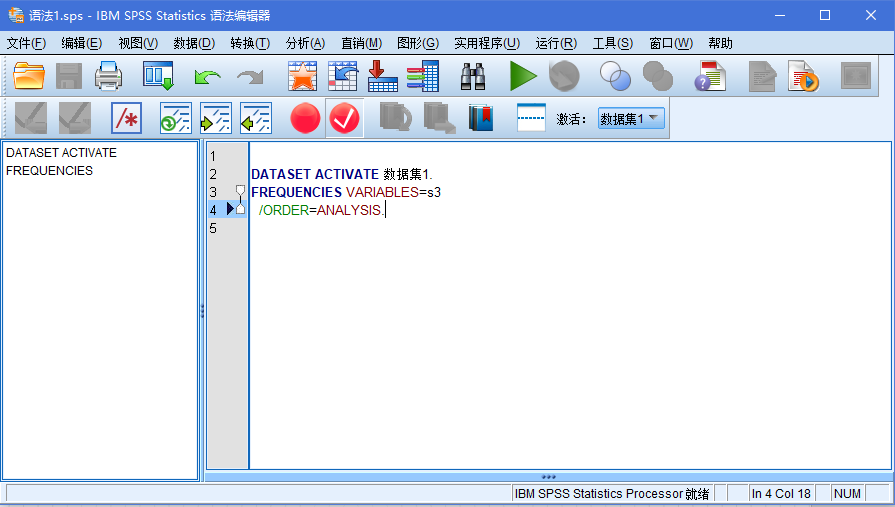
SPSS的模块数量随着版本的不同而有所变化，在18版以前，SPSS Base是必须的，软件的整个框架，基本的数据获取、数据准备等基本功能都被集中在这个模块上，其他模块必须在SPSS Base搭建的平台上才能工作。从18版起，其余模块可以脱离Base单独存在并运行。但对于普通用户而言，仍然是以Base+其余模块的用法最常见。

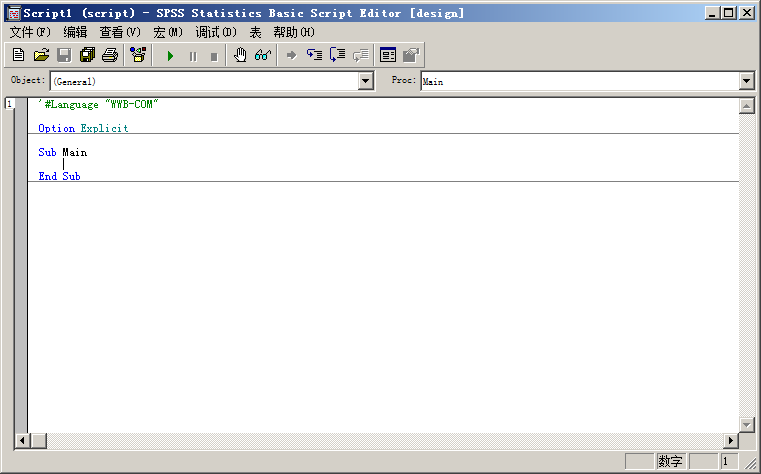


### 【知识点5】SPSS基础操作

SPSS支持**多窗口运行**，运行时最多的窗口种类有4种：数据窗口、输出窗口、语法窗口和脚本窗口。







1）数据窗口：也称为数据编辑器，此窗口类似于Excel窗口，SPSS处理数据的主要工作全在此窗口中进行。又分为2个视图，数据视图用于显示具体的数据，一行代表一个观测个体，一列代表一个属性；变量视图则专门显示有关变量的信息：变量名称、类型、格式等。

单元格编辑器：显示的是当前单元格所在的位置和数值，不能直接输入公式

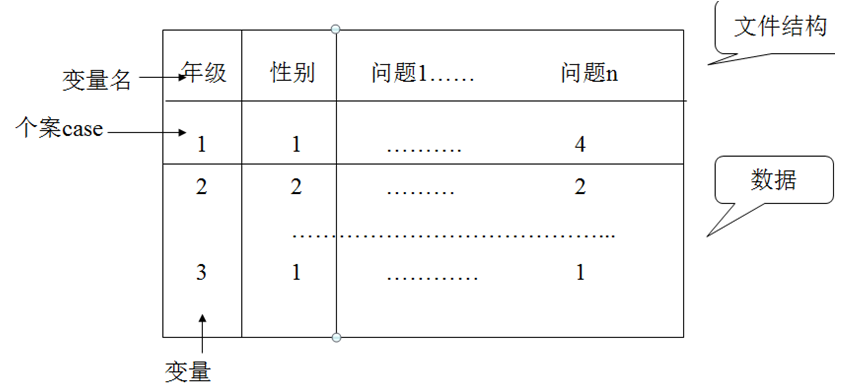
1. 输出窗口：也称为结果查看器，此窗口用于输出分析结果。在窗口中进行的操作非常类似于资源管理器。整个窗口分两个区，左边为目录区，是SPSS分析结果的一个目录，右边是内容区，是与目录一一对应的内容。
2. 语法窗口：也称为语法编辑器，SPSS最大的优势在于菜单操作，该方法是对菜单功能的一个补充。
3. 脚本窗口：SPSS脚本是用Sax Basic语言编写的程序。

### 【知识点6】SPSS的文件菜单

1. 将文件标记为只读：用于锁定当前数据文件为只读，如果之后保存文件，则只能重命名保存。
2. 重新命名数据集：对当前文件的工作名称进行更改，注意是工作名称而不是文件名称。

3、显示数据文件信息：在输出窗口中以表格的形式列出当前文件或指定外部数据文件的信息。

### 【知识点7】SPSS数据文件的特点



SPSS数据文件是一种有结构的数据文件，是二维数据表。

行：case

列：变量

输入数据的步骤：定义变量-设置变量属性-录入数据

### 【知识点8】SPSS数据录入格式要求

* 不同的个案的数据不能在同一条记录中出现，即同一个案例的数据应当独占一行。
* 每一个测量指标/影响因素只能占据一列的位置，即同一个指标的测量数值都应当录入到同一个变量中去。
* 如果出现违反这些原则的例外情况，需对同一个观察对象的某个观察指标进行重复测量；
* 最终的数据集应当能够包含原始数据的所有信息。

### 【知识点9】SPSS的变量名

变量名是变量参与分析的唯一标识，定义变量结构时首先应给出每个变量的变量名，否则系统默认为VAR00001、VAR00002等。

规则： 1）首字符必须为字母或者汉字，后面任意

2）命名长度应少于64个字符**（1个汉字代表3个字符）**

3）不能用下划线\_、句号。、圆点.作为最后一个字符

4）不能与保留字相同：ALL、GT、LE、NE、NOT、OR、TO、WITH等

5）变量名唯一，不区分大小写

### 【知识点10】SPSS变量类型

**数值型**

应用最为广泛，如工资、年龄、成绩等都可定义为数值型

**字符型（分类型变量）**

也是SPSS较常用的数据类型，但由于分析、整理都较困难，建议尽量少用，改为编码录入，如性别用1、2代表男女，用变量值标签加以解释说明

**日期型**

实际上是特殊的数值型变量，尽量少用。日期型数据主要在时间序列分析中比较有用，在较为简单的分析问题中完全可以用数值型变量进行操作：如201610

### 【知识点11】SPSS变量尺度

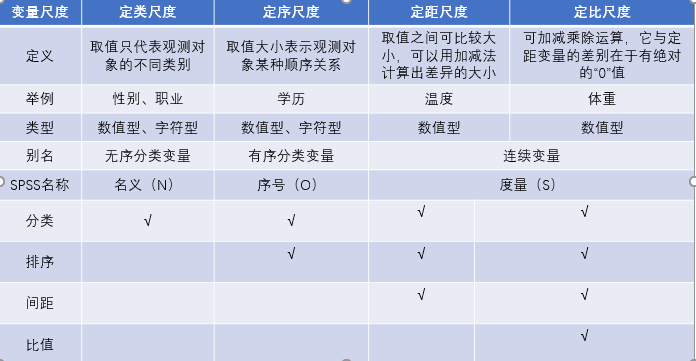
如果只使用变量类型，很多时候并不能准确地说明变量的含义或属性。比如下面这几个变量：

1. 性别：用1代表男，2代表女。在这里1和2只是一个符号，没有任何数字意义，2并不比1大，1也并不比2小。
2. 学历：用1表示初中，2表示高中，3表示本科，1和2虽然也是符号，但这里有一个顺序之分，1就比2的学历低。但究竟低多少，不知道，个级别之间的差距大小无法衡量，更无法进行比较。
3. 年龄：20和21就是有差别的，差1。而且这个距离大小和39和40之间的差距是相等的。

由上可知，变量储存同样都是数值，但数值的具体含义不同，所携带的信息量不同，适用的统计方法也就不同。如果只是一存储类型来说明这个变量的属性，就不能反映上述区别，为此，就有必要给变量增加测量尺度这一属性。

在统计学中，按照对事物描述的精确程度，将所采用的测量尺度从低到高分为4个层次。

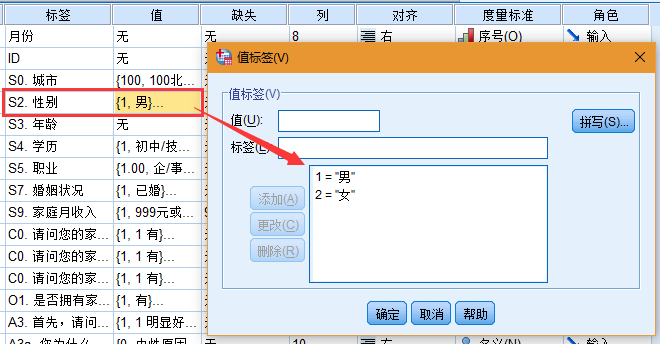
* 定类尺度，是对事物的类别或属性的一种测度，按照事物的某种属性对其进行分类或分组。定类变量的特点是其值仅代表了事物的类别和属性，仅能测定类别差，不能比较各类之间的大小，所以各类之间没有顺序或等级，在SPSS中定类尺度用“名义（N）”来表示。 比如男、女
* 定序尺度，是对事物之间等级或顺序差别的一种测度，可以比较优劣或排序。其计量结果只能排序，不能进行算数运算，SPSS中使用“序号（O）”来表示。 比如初中、高中、大学
* 定距尺度，是对事物类别或次序之间间距的测量。定距变量的特点是其不仅能够将事物区分为不同类型并进行排序，而且可以准确指出类别之间的差距是多少。
* 定比尺度，是能够测量两个测度值之间比值的一种计算尺度，它的测量结果同定距变量一样也表现为数值，与定距变量的差别在于有一个固定的**绝对“零点”**，而定距变量则没有，定距变量中0并不表示没有，仅仅是一个测量值，而定比变量中的0则真正表示没有。比如温度，0℃只是一个普通的温度，而重量则是定比变量，0kg就意味着没有重量可言。**SPSS中将定距和定比合并为一类，统称为“度量（S）/标度”**



### 【知识点12】SPSS的值和标签

标签（Label）：用于定义变量名标签，可用于对变量名的含义进行进一步解释说明，使结果更易于阅读；

值（Values）：用于定义变量值标签，对变量取值含义的解释说明，如：性别，1表示男，2表示女，在数据录入时未设定变量值标签，其他人就很难知晓数据代表的信息。所以变量值标签对于定序变量（如职称）和定类变量（如性别、民族)来说是必不可少的，同样是增强了分析结果的可读性。



### 【知识点13】SPSS的缺失值

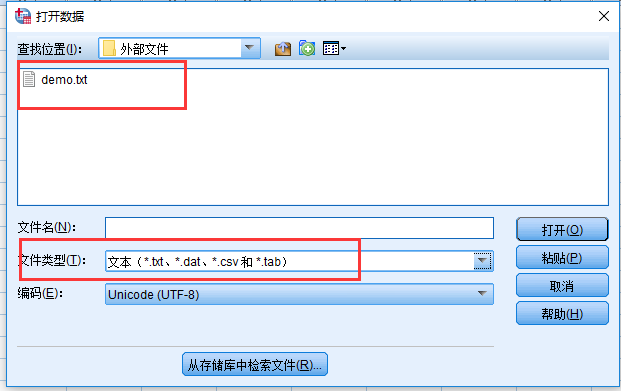
缺失属性是一个重要而且容易被忽略的变量属性，它用于定义变量缺失值。SPSS中缺失值有用户自定义缺失值和系统缺失值两大类。

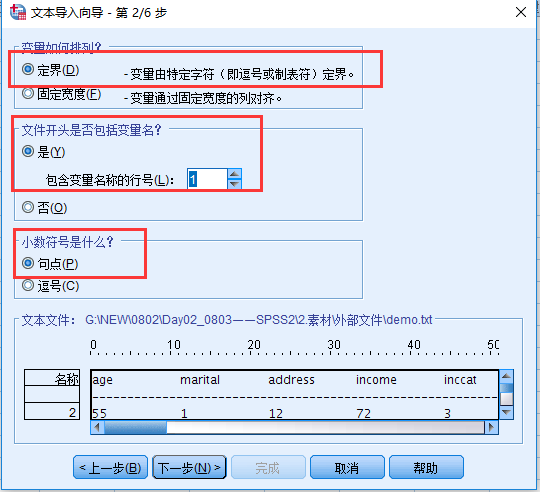
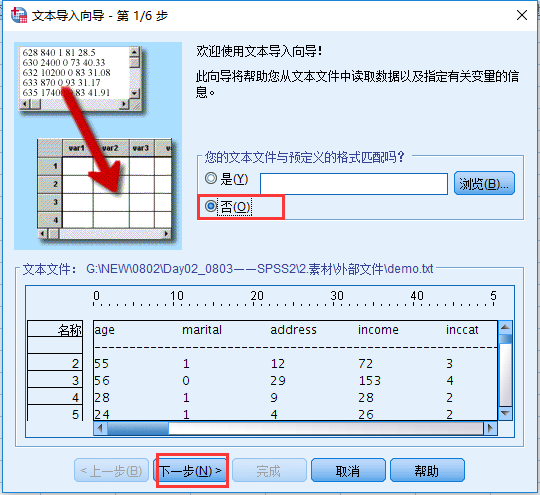
对于数值型变量的数据，系统缺失值用一个圆点“.”表示，而字符型变量默认就是空字符串。

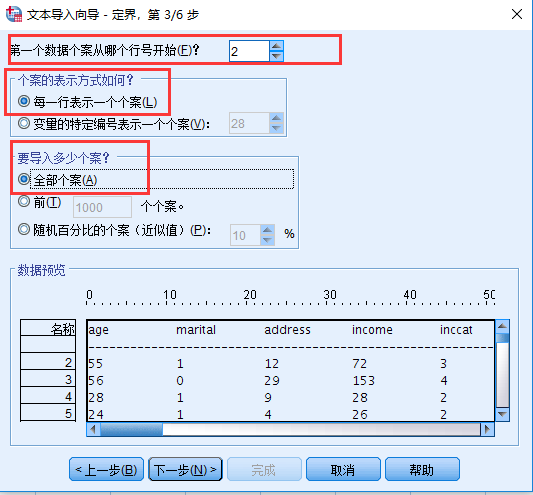
另一类缺失值是用户自定义缺失值，这往往出现在一些设计比较严格的大型调查中，在一些题项中会给出比如不知道/拒答，那么相应的可以用9或者99来表示。



### 【知识点14】导入文本文件

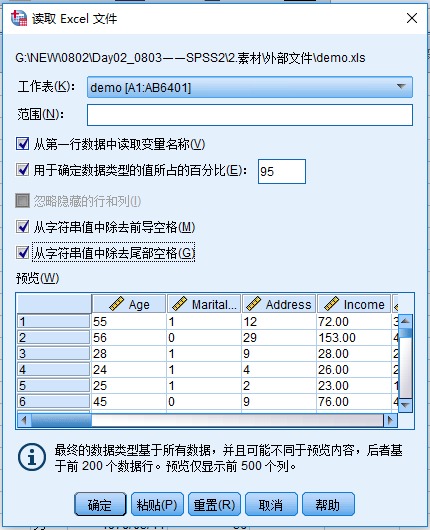








### 【知识点15】导入excel文件



### 【知识点16】手动录入数据

输入姓名、性别

1、单选题：答案只能有一个选项

*当前贵组织机构是否设有面向组织的职业生涯规划系统？*

*A有 B 正在开创 C没有 D曾经有过但已中断*

**编码：**只定义一个变量，Value值1、2、3、4分别代表A、B、C、D四个选项

**录入：**录入选项对应值，如选C则录入3

采用分类法，一个问题有几个答案就把值标签分成几类，将题目设置成变量

2、多选题：答案可以有多个选项，其中又有项数不定多选和项数定多选

***例：****贵处的职业生涯规划系统工作涵盖哪些组群？画钩时请把所有提示考虑在内。*

*A月薪员工 B日薪员工 C钟点工*

**编码：**把每一个相应选项定义为一个变量，每一个变量Value值均如下定义：“0” 未选，“1” 选。

**录入：**被调查者选了的选项录入1、没选录入0，如被调查者选AC，则三个变量分别录入为1、0、1。

这种方法称为二分法

3、项数定多选

***例****：你认为开展保持党员先进性教育活动的最重要的目标是哪三项：*

*1（） 2（） 3（）*

*A、提高党员素质 B、加强基层组织 C、坚持发扬民主*

*D、激发创业热情 E、服务人民群众 F、促进各项工作*

**编码：**定义三个变量分别代表题目中的1、2、3三个括号，三个变量Value值均同样的以对应的选项定义，

即：“1”A，“2”B，“3”C，“4”D，“5”E，“6”F

**录入：**录入的数值1、2、3、4、5、6分别代表ABCDEF，相应录到每个括号对应的变量下。如被调查者三个括号分别选ACF，则在三个变量下分别录入1、3、6。

4、排序题：对选项重要性进行排序

***例****：您购买商品时在①品牌 ②流行 ③质量 ④实用 ⑤价格*

*中对它们的关注程度先后顺序是（请填代号重新排列）第一位 第二位 第三位 第四位 第五位*

**编码：**定义五个变量，分别可以代表第一位到第五位，每个变量的Value都做如下定义：“1” 品牌，“2” 流行，“3” 质量，“4” 实用，“5” 价格

**录入：**录入的数字1、2、3、4、5分别代表五个选项，如被调查者把质量排在第一位则在代表第一位的变量下输入“3”。

5、开放题：这类题目要求被调查者自己填入数值，或者打分

***例****：你的年龄（实岁）：\_\_\_\_\_\_*

**编码：**一个变量，不定义Value值

**录入：**录入被调查者实际填入的数值。

**注意：**对于开放性问题，如果可能的话可以按照含义相似的答案进行编码，转换成为封闭式选项进行分析。如果答案内容较为丰富、不容易归类的，应对这类问题直接做定性分析。

### 【知识点17】变量赋值

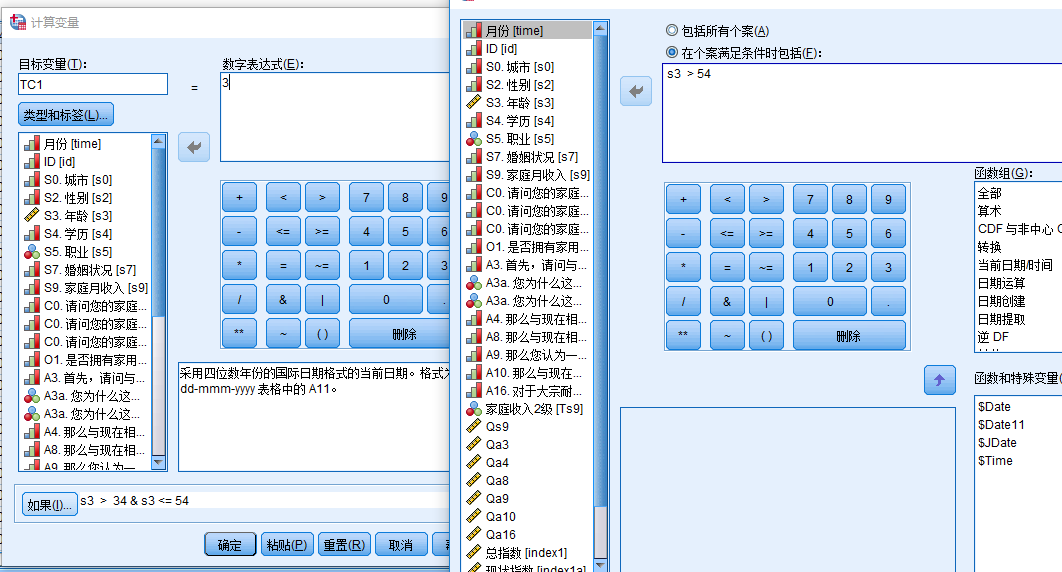
条件表达式：**~=**表示不等于

逻辑表达式：&表示and、|表示or、~表示not

位置：转换——计算变量

例子：CCSS项目中的受访者年龄在18到65岁，现在分成<=34（1）、35到54（2）、>=55（3）三组

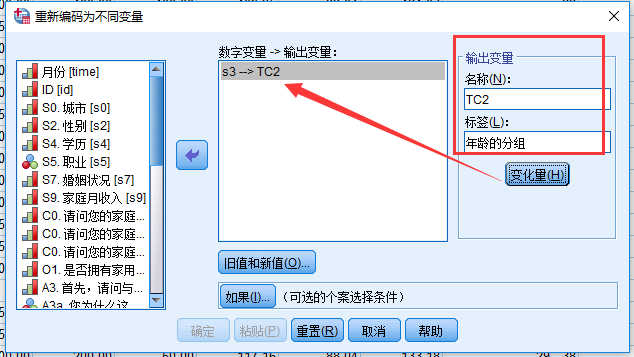




### 【知识点18】重新编码为（相同）不同变量

位置：转换——重新编码为不同变量

例子：CCSS项目中的受访者年龄在18到65岁，现在分成18到34、35到54、55到65三组





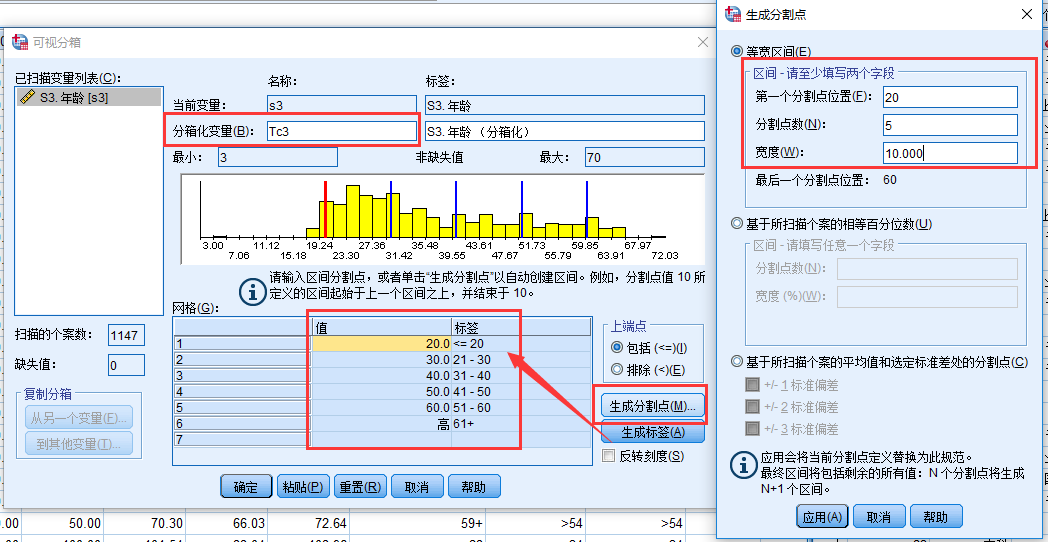
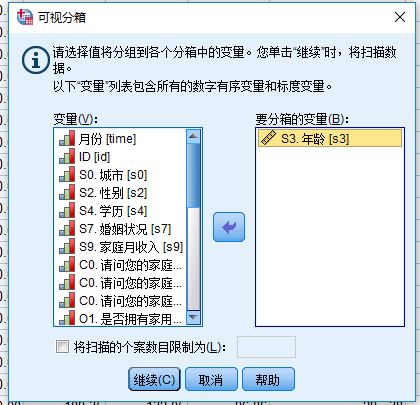
### 【知识点19】数值变量的离散化分段

**目的：**根据分析目的将数值型数据进行等距或非等距分组，该过程也称为分箱，一般用于查看分布情况

**操作步骤**：转换-可视离散化(可视分箱)

**注意：**分割点各参数的设置及生成标签

数据：用户明细.sav



### 【知识点20】变量的自动重编码

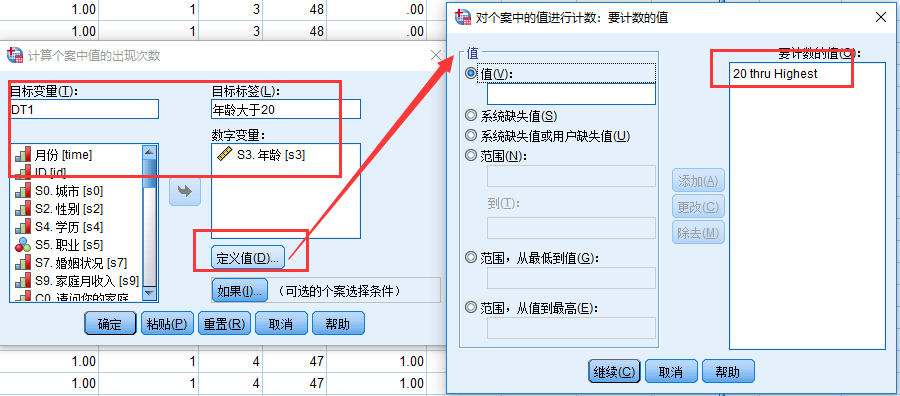
位置：转换——自动重新编码

例子：将年龄自动重编码

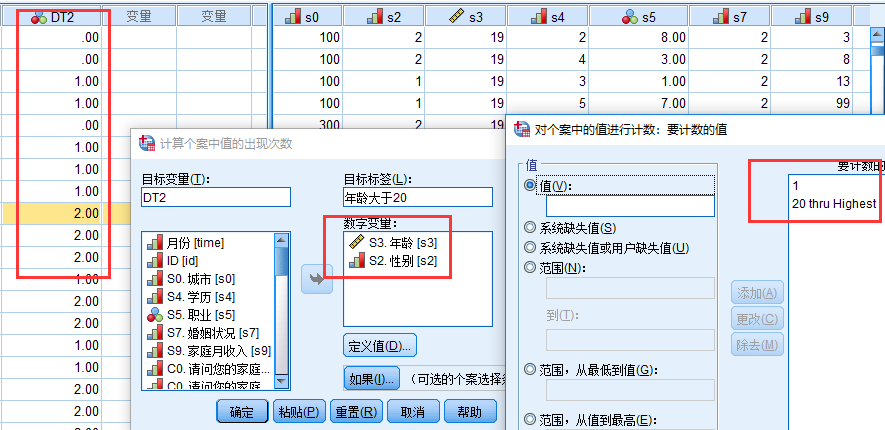
### 【知识点21】计算个案内值的出现次数

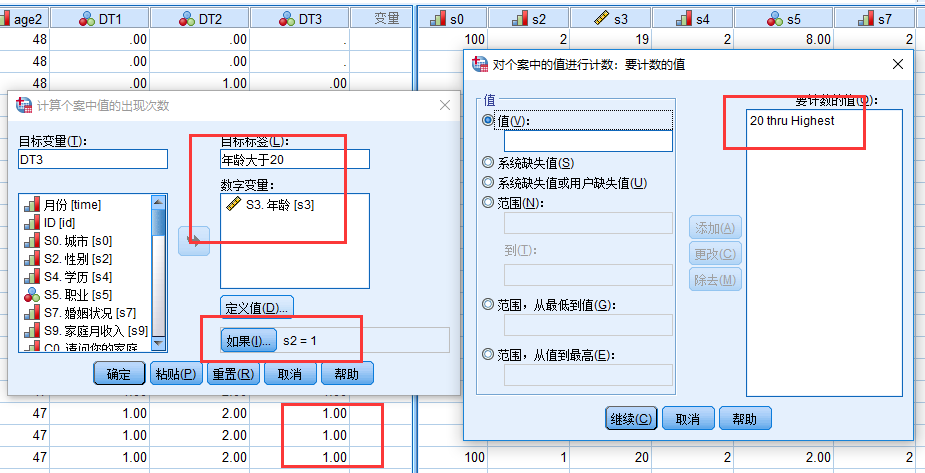
位置：转换——对个案内的值计数

对年龄大于20的进行标记



对年龄大于20并且性别为男性的进行标记





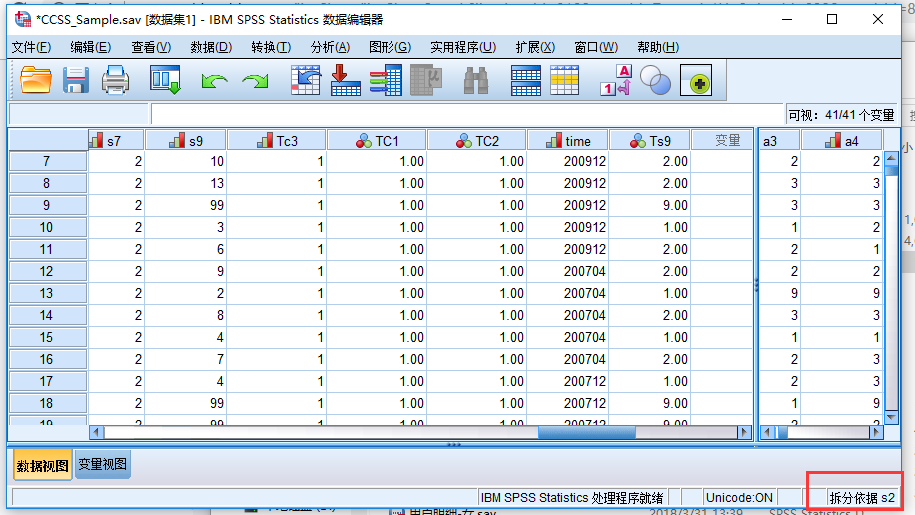
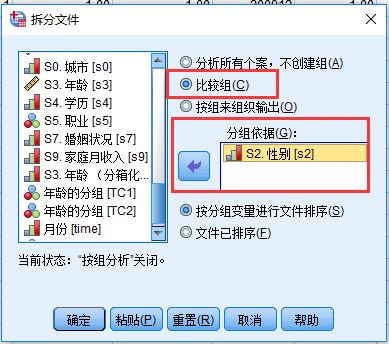
### 【知识点22】记录排序

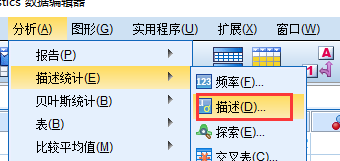
位置：数据——排序个案

目的：便于查找、发现错误、发现缺失值

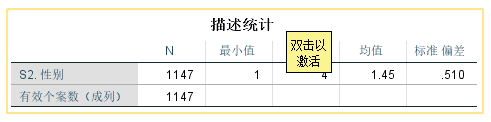
### 【知识点23】记录拆分

位置：数据——拆分文件





取消拆分



### 【知识点24】记录筛选

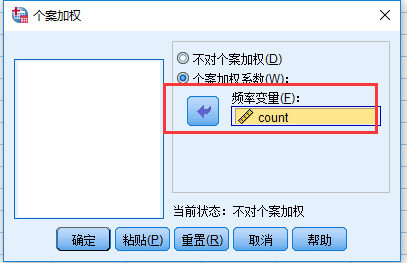
位置：数据——选择个案

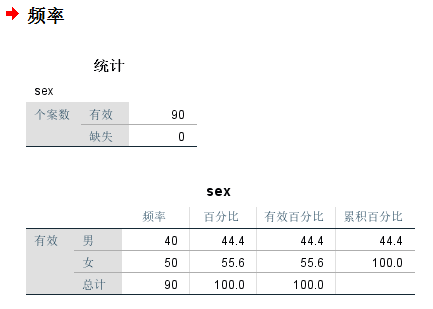
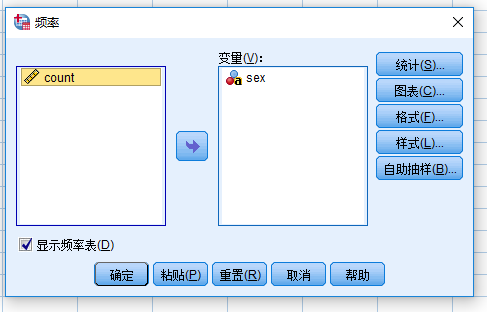
例子：筛选出年龄小于30岁的个案

### 【知识点25】记录加权

位置：数据——个案加权

注意：设置权重变量



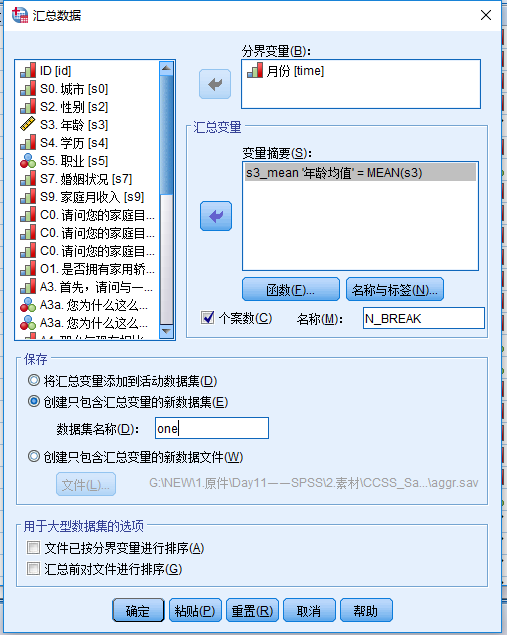


### 【知识点26】分类汇总

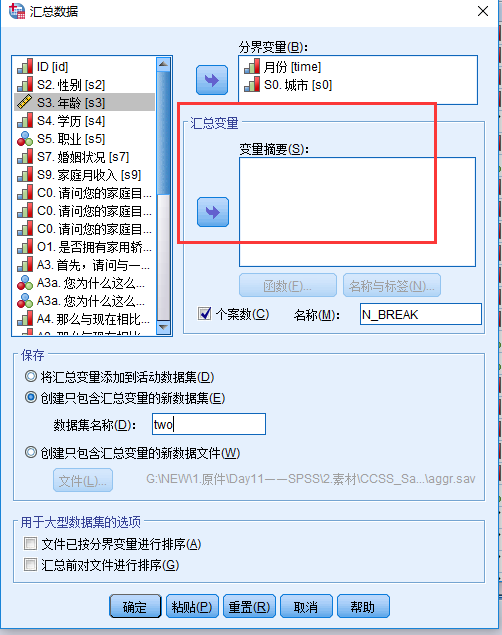
位置：数据——分类汇总

分类汇总：按指定的分类变量对个案进行分组，并按分组对变量求指定的描述统计量，结果可以存入新数据文件中，也可以替代当前数据文件。

查看不同月份的平均年龄



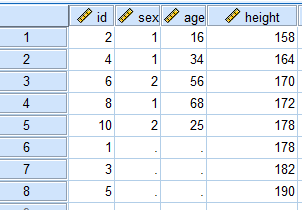
例子：对不同月份、不同城市进行分组



### 【知识点27】多个数据文件的合并

纵向合并：从外部数据文件中增加记录到当前数据文件（追加查询）

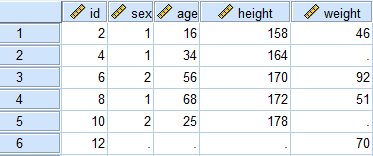
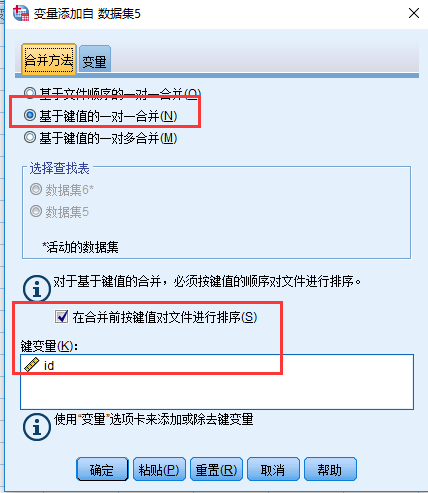
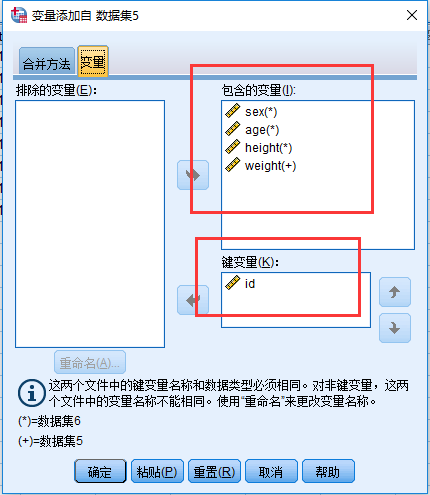
位置：数据——合并文件——添加个案



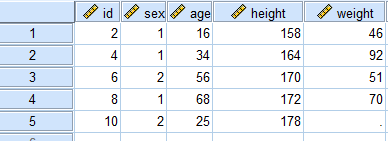
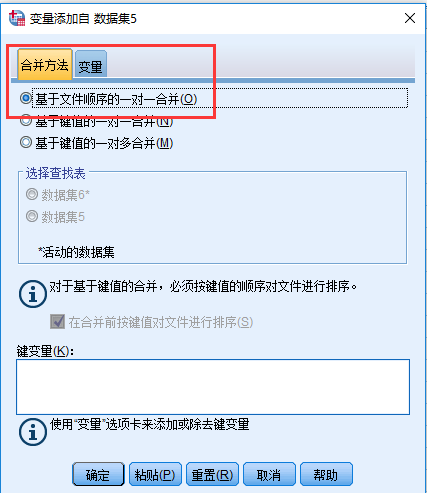
横向合并：从外部数据文件增加变量到当前数据文件

位置：数据——合并文件——添加变量

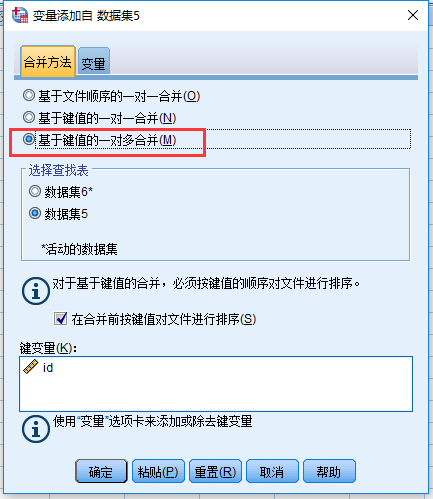
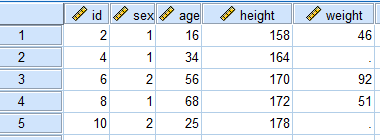
基于键值的一对一合并：类似与power query的合并的查询的完全外部（也就是a和c两张表的所有数据）



基于文件顺序的一对一合并



基于键值的一对多合并：类似于power query的合并查询的左外部（也就是左边这张表的所有数据a）

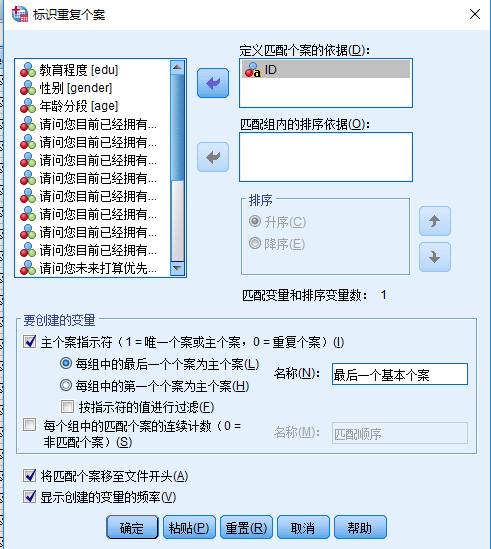
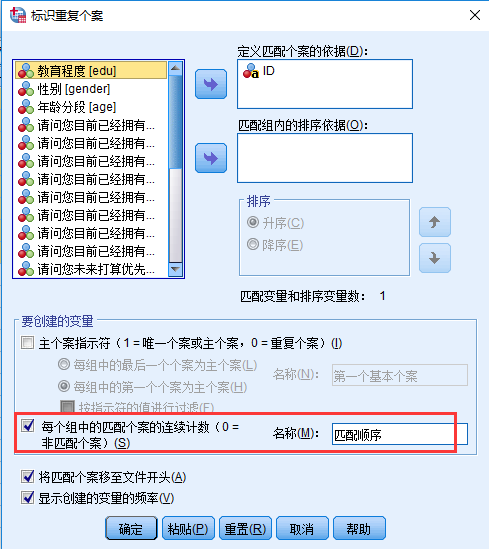
### 【知识点28】标识重复个案

位置：数据——标识重复个案

数据源：1.问卷录入数据(整理后).sav

通过数据——选择个案删除重复个案

第二种方式：匹配顺序：就是对所有重复的个案计数从1开始依次往后计数，不重复的个案直接标记为0



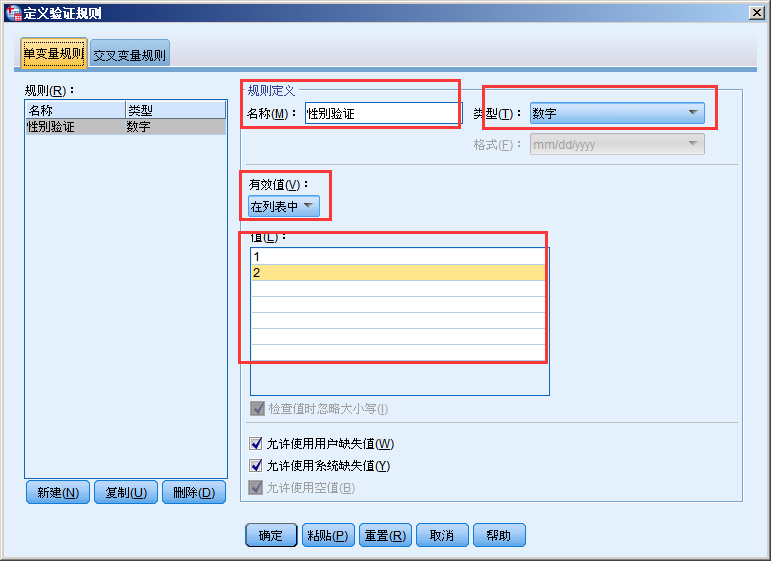
### 【知识点29】数据验证

判断数据有效性

通过数据验证可以找出数据集中违反验证规则的数据，基本步骤是先定义验证的规则（单变量或交叉变量），再对指定的变量或整个数据集用设定好的规则进行检验。

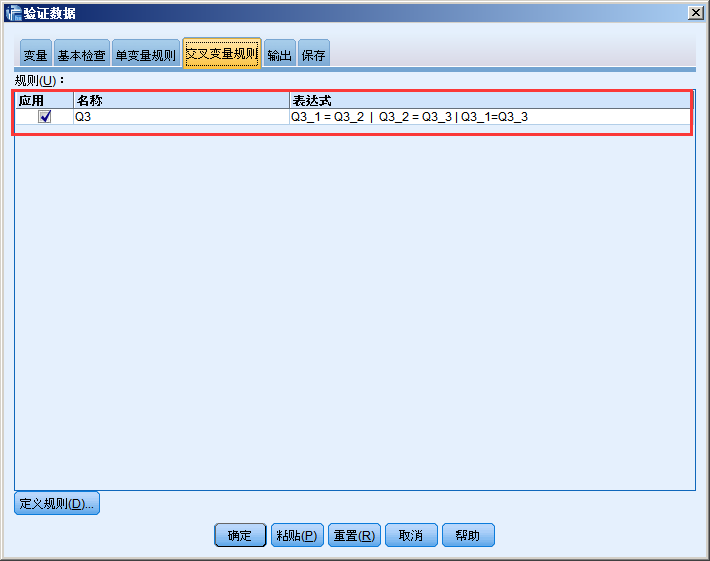
数据-数据验证

定义规则



验证规则





## 自主学习作业讲解

无

## 课程总结

* SPSS简介
* SPSS产品
* SPSS特点
* SPSS模块
* SPSS基础操作
* SPSS的文件菜单
* SPSS数据文件的特点
* SPSS数据录入格式要求
* SPSS的变量名
* SPSS变量类型
* SPSS变量尺度
* SPSS的值和标签
* SPSS的缺失值
* 导入文本文件
* 导入excel文件
* 手动录入数据
* 变量赋值
* 重新编码为（相同）不同变量
* 数值变量的离散化分段
* 计算个案内值的出现次数
* 记录排序
* 记录拆分
* 记录筛选
* 记录加权
* 分类汇总
* 多个数据文件的合并
* 标识重复个案
* 数据验证

## 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