1.2.3 SPSS 的操作方式

1. 统计软件的常见操作方式

初学者对 SPSS 存在的一个广泛误解是 SPSS 只使用菜单对话框方式来操作。实际上,在经历了几十年的发展之后,现今任何一个成熟的统计软件都会提供从初学者到专家各个层面所需要使用的操作方式。具体而言,统计软件常见的操作方式有如下几种。

(1) 命令行方式:即用户一条一条地提交命令,软件系统直接对命令进行解释执行,用户再根据执行结果提交下一条命令。这是出现最早的一种操作方式,也是目前大型统计软件中 Stata 的主要操作方式。

(2) 程序方式:由于命令行方式无法实现一些复杂功能,因此随后就出现了将命令组合成程序,用户批量提交,系统按程序要求执行,批量输出结果的程序方式。程序方式不仅可以提高运行效率,还可以利用程序结构中的分支、循环等语句来实现更复杂的统计功能。目前统计软件中的 SAS 就是以程序方式作为主要执行方式的。

(3) 菜单对话框方式:由于程序方式需要用户首先学习编程语法规则,对于初学者来说学习门槛较高,因此 SPSS 针对初学者的需求在全球首家推出了以菜单对话框为主要操作方式的软件版本,在随后的几十年里这种操作方式成为 SPSS 的主要操作方式。这一方式随着 Windows 系统的普及而得到了极大的发展,也已经成为了最受各阶层统计软件用户欢迎的方式。

需要指出的是,各种大型统计软件实际上都支持上述 3 类操作方式,只是有其首选方式和最佳操作方式的区别而已。例如,上面提到的 Stata 和 SAS 都支持菜单对话框方式,而 SPSS 也支持程序方式。

2. SPSS 对各种操作方式的支持和扩展

具体而言,作为全球知名的统计软件,SPSS 对上述几种操作方式都是支持的,还进一步提供了一些扩展,其重点在于提高软件的易用性。

(1) 菜单对话框方式:用户可以自定义对话框,以便将工作中一些常用的程序直接以对话框方式实现,其实质就是对软件进行二次开发。对菜单对话框操作方式基本规则的介绍可参见下文。

(2) 命令行/程序方式:作为最易学、易用的统计软件,SPSS 对程序方式做了很大的改进。首先可以利用对话框中的"粘贴"按钮自动生成程序,其次可以通过宏、Include 命令等方法使得已有代码段得到充分利用;最后,SPSS 还提供了程序全自动执行的"生产工作(Production Facility)"方式,进一步简化了操作。对程序方式操作细节的详细介绍参见本书第 5 章。

1.2.4 SPSS 对话框操作基本规范

SPSS 的对话框都遵循统一的操作规范,因此这里首先介绍一下相应的规范,以便于后续学习。

1. 对话框元素构成

这里以"频率"对话框为例,选择"分析"→"描述统计"→"频率"菜单项,打开"频率"对话框,如图 1.2 所示。

（2）字符型（String）：字符型也是SPSS较常用的数据类型，字符型数据的默认显示宽度为8个字符位，它区分大小写字母，并且不能进行数学运算。字符型数据在SPSS的数据处理过程（如在计算生成新变量时）中是用一对引号引起来的。需要注意的是，在输入数据时不应输入引号，否则，双引号将会作为字符型数据的一部分。

（3）日期型（Date）：这种类型的数据是用来表示日期或时间的。日期型数据的显示格式有很多，SPSS在对话框右侧会以列表框的方式列出各种显示格式以供用户选择。如果此处选择的是mm/dd/yy或类似的两位数年份记录方式，则需要在系统选项的“数据”选项卡中确定具体的世纪范围，目前系统默认为1941—2040年区间。

事实上，SPSS中的日期型变量存储的是该时间与1582年10月14日零点相差的秒数，如1582年10月15日存储的就是60×60×24=86 400，将变量类型变换为数值型就可以看到。但是这里只能存储正数，即1582年10月14日及更早的时间在SPSS中是无效的。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普通数值型数据来代替。

2. 变量的测量尺度

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和属性。比如CCSS数据中的以下几个变量。

（1）变量S2“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）变量S4“学历”：用1表示“初中”，2表示“高中”，3表示“本科”等，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。但是究竟低多少，是本科和高中的差距更大，还是高中和初中的差距更大，不知道，各级别之间的差距大小无法衡量，更无法进行比较。

（3）变量S3“年龄”：20和21就是有区别的，差1。而且这个差距大小，和39与40之间的差距是相等的，都是1，也等于50和55之间差距的1/5。

由上可知，上述3个变量的存储类型同样都是数值型，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。如果只以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、定序尺度、定距尺度和定比尺度。在这4种测量尺度之间，按照信息量的高低，可将高层次测量尺度的测量结果转换为低层次测量尺度的测量结果，但这样会损失一部分信息，但不能将低层次的测量尺度转换为高层次测量尺度的结果，这样可能会引入错误的信息。

1）定类尺度

定类尺度（Nominal Measurement）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，如变量S0“城市”就是一个定类尺度的变量。对于定类尺度的变量只能计算频数和频率，如在所有个案中，北京有多少人，占总人数的百分率是多少等。对于S2“性别”这种两分类变量，一般仍然将其归为定类尺度变量。但是两分类变量较为特殊，即使将其归为其他类型，一般也不会影响后续分析。

在SPSS中使用度量标准（Measure）属性对变量的测量尺度进行定义，其中定类尺度变量用“名义（N）”来表示。能使用的定类尺度的数据可以是数值型变量，也可以是字符型变量。使

表 2.1　SPSS 可以直接打开的数据类型

数据标识	数据类型
SPSS Statistics (*. sav)	SPSS 各版本的数据文件
SPSS/PC + (*. sys)	SPSS/PC + 版本的数据文件
Systat(*. syd, *. sys)	Systat 数据文件
便携(*. por)	SPSS 便携格式的数据文件
Excel(*. xls, *. xlsx, *. xlsm)	Excel 各版本的数据文件
Lotus(*. w*)	Lotus 各版本的数据文件
SYLK(*. slk)	以 SYLK(符号链接)格式保存的数据文件
dBASE(*. dbf)	dBASE 系列数据文件(从 dBASE Ⅱ ~ Ⅳ)
SAS(*. sas7bdat, *. sd7,…)	SAS 各版本的数据文件
Stata(*. dat)	Stata 4 ~ 8 版的数据文件
文本格式(*. txt, *. dat)	纯文本格式的数据文件

和老版本的菜单相比,文件类型列表框做了大幅的合并,显得更为简洁。在其中选择所需的文件类型,然后选中需要打开的文件,SPSS 就会按照要求打开相应的数据文件,并自动转换为 SPSS 格式。

2. 操作实例

下面以 SPSS 自带的文件 demo. xls 为例说明 SPSS 如何直接读取这个文件,该文件位于 SPSS 安装目录下的 Samples 子目录中。首先在 Excel 中打开 demo. xls,了解一下这个文件的结构,重点需要了解这样几项内容:第一,该文件中包含几个数据表,具体应当打开哪个表;第二,如果不需要该表的所有数据,而只需读入一部分,这时需要了解要读入数据的精确位置,如单元格 A2:F5;第三,此部分数据的第 1 行是否是变量名。从这个文件中可以看出,第 1 行是变量名,该文件只有一个表,要读取的是该表单中的全部数据。

第一步,在打开文件对话框中,选择路径(此例中为 Samples\English),选择文件类型 Excel (. xls),文件列表中出现所有的 Excel 文件,单击文件 demo. xls;第二步,打开如图 2.12 所示的对话框:在“工作表”下拉列表框中选择一个表;在“范围”文本框中指定读取的数据的具体位置,用单元格的起(左上角单元格名称,如 A2)止(右下角单元格名称,如 F5)位置来表示,中间用冒号“:”隔开;上方的复选框用于确定单元格范围的第 1 行是否为变量名。指定完毕,单击“确定”按钮,数据就会被顺利地读入 SPSS 中。

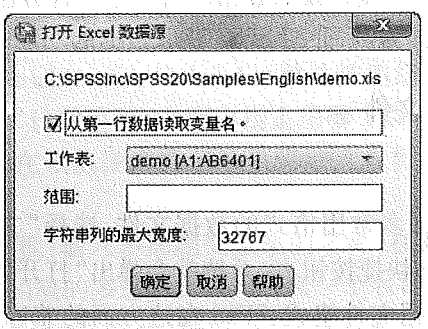


图 2.12　“打开 Excel 数据源”对话框

3. 将 Excel 或 Word 中的数据直接导入 SPSS

对于 Excel 数据文件而言，如果在 Excel 中已经打开原数据文件，并且数据量较少，可以直接用复制、粘贴的方法将数据引入 SPSS 中。先在 Excel 中选中所有的数据（不包括变量名），然后执行“复制”命令；然后切换到 SPSS，最好使行 1 列 1 单元格成为当前单元格，然后执行“粘贴”命令，数据就会全部转入 SPSS 中，再定义相应的变量即可。

如果数据中含有文本，则不能直接粘贴；否则会丢失数据。这是因为 SPSS 默认的数据格式均为数值型，这样将文本粘贴过来就会变为缺失值。解决办法是先在 SPSS 中设定好相应的变量列表，包括数值型、字符串型这些属性，然后再对相应的列进行粘贴，此时字符串数据就不会丢失了。

对于 Word 文档中的数据表格，其操作方式和 XLS 文件基本相同，粘贴后原来的单元格会自动对应为 SPSS 中的一个单元格。

4. 快速改变变量排列次序

在数据视图中选中列首的相应变量名，松开左键后再按下左键不放，就可以将该列数据拖动到任何希望的地方去了。选择时可以选中连续的多个变量，此时这些变量会同时发生变化；但如果选中不连续的多个变量，拖动时只对居中的一个起作用。

该操作也可以在变量视图中进行，此时应当选中变量的相应行号，其余操作相同。

5. 快速定位记录

记录的快速定位在大型数据集操作中非常有用，具体可以分成以下两种情况。

（1）快速定位到第 N 条记录，此时可选择“编辑”→“转至个案”菜单项，或者直接单击工具栏上的按钮，在打开的对话框中输入相应的记录号，单击“确定”按钮后即可。

（2）定位到变量值等于某个取值的记录（如 ID = 34980），此时需要先使相应的变量成为当前列，然后单击按钮，在打开的“查找”对话框中输入相应的数值，单击“确定”按钮后系统就会查找到符合条件的第 1 条记录，单击“查找下一个”按钮则会继续查找第 2 条记录。当然，如果事先排好序，则查找一次就可以了。

6. 利用排序功能快速查找异常值、极端值

对于异常值、极端值的发现，标准的做法应当是做出频数表看看有无异常值，但这样做过于麻烦，而且往往无法马上知道是哪一条记录出错了。其实最简单的做法是在数据视图中选中列首的相应变量名，然后右击，根据需要选择快捷菜单下方的“升序排列”或“降序排列”菜单项，相应的最小值（或缺失值）、最大值就会成为第 1 条记录。现在数据中有无异常值、极端值即可一目了然。

7. 利用变量值标签检查录入错误

前面曾经提到过对于单选题，最好采用“数值代码 + 值标签”的方式进行录入。这里再详细解释一下用法：实际上除了注释变量外，绝大多数字符型变量都只有少数的几种取值，因此可以将这些变量一律按照数值型变量来设置，录入时只需要输入编制的代码 1，2，3，…，然后利用变量值标签将实际含义一一写入标签中，这样可以大大加快速度。

下面来进行最重要的一步：在菜单栏上选择“视图”→“值标签”菜单项，或者直接单击工具栏上的按钮，在该按钮被按下后，数据编辑器中所有设定了值标签的变量值均会切换成相应的值标签。该按钮弹起后，则仍然按照录入的数值来显示，这不仅仅只是一个好看的问题，当单

3. 将 Excel 或 Word 中的数据直接导入 SPSS

对于 Excel 数据文件而言，如果在 Excel 中已经打开原数据文件，并且数据量较少，可以直接用复制、粘贴的方法将数据引入 SPSS 中。先在 Excel 中选中所有的数据（不包括变量名），然后执行“复制”命令；然后切换到 SPSS，最好使行 1 列 1 单元格成为当前单元格，然后执行“粘贴”命令，数据就会全部转入 SPSS 中，再定义相应的变量即可。

如果数据中含有文本，则不能直接粘贴；否则会丢失数据。这是因为 SPSS 默认的数据格式均为数值型，这样将文本粘贴过来就会变为缺失值。解决办法是先在 SPSS 中设定好相应的变量列表，包括数值型、字符串型这些属性，然后再对相应的列进行粘贴，此时字符串数据就不会丢失了。

对于 Word 文档中的数据表格，其操作方式和 XLS 文件基本相同，粘贴后原来的单元格会自动对应为 SPSS 中的一个单元格。

4. 快速改变变量排列次序

在数据视图中选中列首的相应变量名，松开左键后再按下左键不放，就可以将该列数据拖动到任何希望的地方去了。选择时可以选中连续的多个变量，此时这些变量会同时发生变化；如果选中不连续的多个变量，拖动时只对居中的一个起作用。

该操作也可以在变量视图中进行，此时应当选中变量的相应行号，其余操作相同。

5. 快速定位记录

记录的快速定位在大型数据集操作中非常有用，具体可以分成以下两种情况。

（1）快速定位到第 N 条记录，此时可选择“编辑”→“转至个案”菜单项，或者直接单击工具栏上的按钮，在打开的对话框中输入相应的记录号，单击“确定”按钮后即可。

（2）定位到变量值等于某个取值的记录（如 ID = 34980），此时需要先使相应的变量成为当前列，然后单击按钮，在打开的“查找”对话框中输入相应的数值，单击“确定”按钮后系统就会查找到符合条件的第 1 条记录，单击“查找下一个”按钮则会继续查找第 2 条记录。当然，如果事先排好序，则查找一次就可以了。

6. 利用排序功能快速查找异常值、极端值

对于异常值、极端值的发现，标准的做法应当是做出频数表看看有无异常值，但这样做过于麻烦，而且往往无法马上知道是哪一条记录出错了。其实最简单的做法是在数据视图中选中列首的相应变量名，然后右击，根据需要选择快捷菜单下方的“升序排列”或“降序排列”菜单项，相应的最小值（或缺失值）、最大值就会成为第 1 条记录。现在数据中有无异常值、极端值即可一目了然。

7. 利用变量值标签检查录入错误

前面曾经提到过对于单选题，最好采用“数值代码 + 值标签”的方式进行录入。这里再详细解释一下用法：实际上除了注释变量外，绝大多数字符型变量都只有少数的几种取值，因此可以将这些变量一律按照数值型变量来设置，录入时只需要输入编制的代码 1，2，3，…，然后利用变量值标签将实际含义一一写入标签中，这样可以大大加快速度。

下面来进行最重要的一步：在菜单栏上选择“视图”→“值标签”菜单项，或者直接单击工具栏上的按钮，在该按钮被按下后，数据编辑器中所有设定了值标签的变量值均会切换成相应的值标签。该按钮弹起后，则仍然按照录入的数值来显示，这不仅仅只是一个好看的问题，当单

3.3　连续性变量的离散化

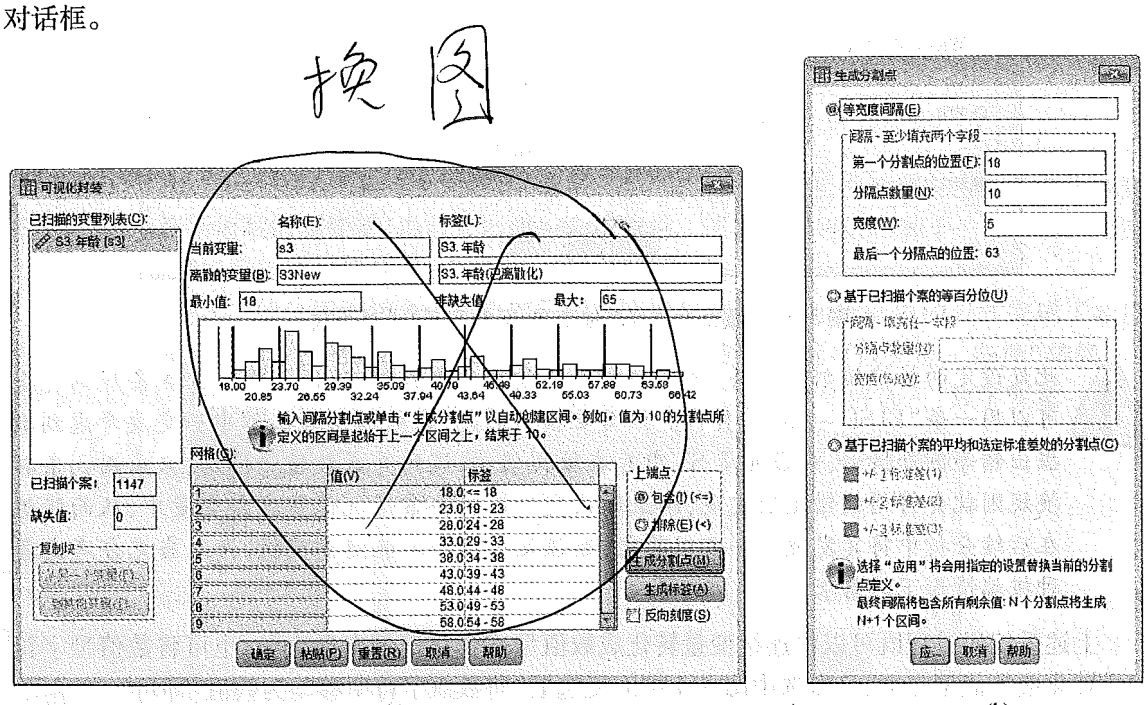
重编码过程提供了精确分组的功能，但是如果希望进行的分组是较有规律的，比如说等距分组，或者等样本量分组，使用重编码过程进行操作就显得非常麻烦，且可视化程度不高，此时可以考虑使用可视化过程进行分段。SPSS 中提供了两种可视化分段过程，分别为需要用户自行判断设定的可视离散化过程，以及基本全自动的最优离散化过程。

3.3.1　可视离散化过程

可视离散化过程用于在可视界面下将连续变量进行分段，在该过程中可以使用百分位数、标准差范围或者等间距方式将连续变量划分为若干组段，并采用图形化操作的方式，非常直观好用。

1. 对话框

选择“转换”→“可视离散化”菜单项之后，打开的对话框要求用户选择希望进行离散化的变量，选择完毕后单击“继续”按钮，则系统会对相应的变量进行数值扫描，并打开如图 3.5 所示的对话框。



(a)　　(b)

图 3.5　“可视化封装”对话框和“生成分割点”对话框

(1) 已扫描的变量列表：列出了在前一个对话框中所选择的所有变量，更改此处的变量选择，则对话框中所有其余部分的内容均会按照所选中的变量状况进行更新。

(2) 左下侧“复制块”框组：当选择了多个变量且其中部分变量已设定完离散化选项时可用，可以将设定好的属性复制“到其他变量”，也可以“从另一个变量”（即已设定好的变量）读取

相应的设定。

(3) 上部变量属性:列出新老变量的名称和标签,注意其中新变量名称是必填的;否则离散化完毕后不会生成任何新变量。

(4) 中部直方图:扫描完原变量取值情况后在此处绘制该变量的直方图,如果已设定完毕分割点,也会一并显示。

(5) 下部数值标签网格:在本网格处显示所设定的分割点数值位置和相应的标签。

(6) 右下侧"上端点"框组:用于设定端点是否被包括在上侧区间内。

(7) "生成分割点"按钮:单击后打开如图 3.5(b)所示的对话框,其中可以选择使用等间距(Equal Width Inter)、等比例(等样本量,Equal Percentiles based on Scanned Cases)或者按照指定的标准差范围(Cutpoints at Mean and Selected Standard Deviations based on Scanned Cases)3 种方式进行分段,其中第 3 种方式显然可以用来在数据分析或质量控制中筛选异常值。

(8) "生成标签"按钮:在分割点数值设定完毕后,单击该按钮可以自动生成相应的值标签。

(9) "反向刻度"复选框:在默认情况下,新的离散化变量的值是从 1 到 n 的升序整数。反向刻度会使得这些值成为从 n 到 1 的降序整数。

2. 实例分析

例 3.2 将 S3 年龄变量分为 10 组,要求等间距。

本例实际上是要求对连续变量进行统计描述中的直方图分组,由于已知年龄范围为 18 ~ 65 岁,全距为 48,因此在分为 10 组的情况下,组距为 5 即可覆盖全部取值范围。当然组数、组距和第一组段下限三者是相互联系的,在对话框中一般只需要定义其中两者即可自动确定第 3 个因素的取值。

(1) 选择"转换"→"可视离散化"菜单项,将 S3 年龄选入"要离散的变量"列表框中,单击"继续"按钮进入主对话框。

(2) 单击"生成分割点"按钮,设定分割点数量为 10,宽度为 5,可见系统会自动填充第一个分割点的位置为 18,单击"应用"返回到主对话框。

(3) 此时会看到下部数值标签网格的"值"列已被自动填充,单击"生成标签"按钮,使标签列也得到自动填充。

选中右侧"上端点"框组中的"排除"复选框,然后

(4) 将离散的变量的名称设定为 S3New,单击"确定"按钮,系统会提示"封装规范将创建一个变量",确认后就会在数据集中生成新变量 S3New。

如果注意一下结果窗口中的 LOG 输出就会发现,可视化分段过程实际上运行的是记录重编码所对应的 Recode 过程,也就是说,两者在代码级别上实际是一回事情,只不过可视化分段过程在对话框界面上进行了进一步的开发而已。

3.3.2 最优离散化过程

"最优离散化"过程是对前述可视化离散过程的进一步自动化,根据某些作为"关键指示变量"的分类变量,将原有的一个或多个连续性变量按照该分类变量类间差异最大化的优化原则离散化为分类变量,然后就可以使用离散化变量而非原始数据值进行后续的分析了。

初学者可能对该过程的用途感到疑惑,实际上这主要和建模分析有关:当模型中的因变量为分类变量时,在分析中往往会对自变量进行离散化(分类化),此时就可以使用该过程。如果最

(6)“保存”框组:设定汇总结果的具体输出方式,可以将汇总后的结果直接加入当前数据文件中,也可以定义一个新文件以存储汇总的结果,或者用汇总的结果替换原来的数据。

2. 实例分析

例 4.1　按 time 月份和 s0 城市对 CCSS 案例数据中的变量 index1 进行均数汇总,并将结果输出到新数据文件 Sum_index1 中。

本例的分组变量不止一个,此时第一个指定的分类变量为主分类变量,其他的依次为第 2、第 3 分类变量,且汇总数据文件的记录数等于各分类变量类别数的乘积,因此本例的汇总数据文件中会有 4×3=12 条记录。

按照图 4.9 所示的方式进行对话框设定,操作完毕后 SPSS 会建立一个新数据文件,其中存储的就是相应的汇总结果,如图 4.10 所示。

	time	s0	index1_mean	N_BREAK
1	200704	100北京	100.05	100
2	200704	200上海	97.79	100
3	200704	300广州	97.16	100
4	200712	100北京	97.13	101
5	200712	200上海	92.10	101
6	200712	300广州	93.19	102
7	200812	100北京	91.97	102
8	200812	200上海	87.68	102
9	200812	300广州	91.70	100
10	200912	100北京	102.58	75
11	200912	200上海	102.56	84
12	200912	300广州	100.86	80

图 4.10　数据 Sum_index1 的内容

4.2　数据文件的重组与转置

用到的一个功能。数据录入的格式未必能一步到
要求改变数据的排列格式,数据重组是一个图形化

4. 种不

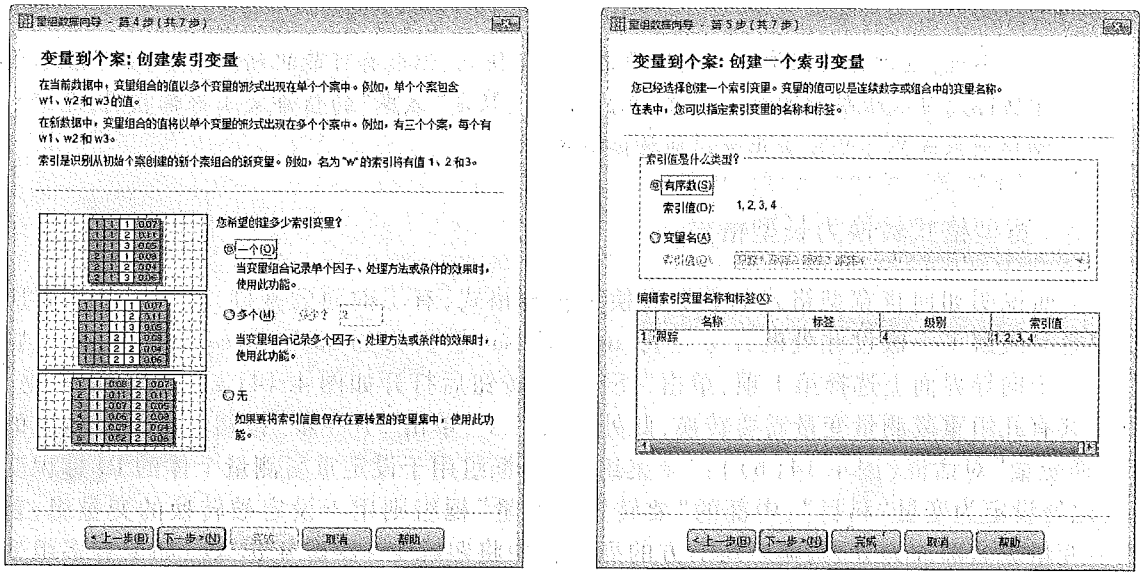
所讲
函数,分别
用的第一组为
录分别计算汇总
量,结果可以
作中经常遇到的
况,这时就需要首先
本质上就是一个数据的
生的汇总变量的名称和

1. 界面简介

分类汇总过程涉及的对话
条数。

据的两种不同的排列方式,由于重复测量模型可以
求进行长型格式和宽型格式之间的互转实际上是

anxiety2. sav 来说明这两种数据排列格式的特点。
受治疗后的 4 个时间点的精神状态评分,其中“科
的 ID 号,“评分”(Score)为评估的分数,“跟踪”
)和紧张(Tension)记录了病人在治疗前是否焦虑和



(a)　　(b)

图 4.15　数据重组向导的第 4、5 个界面

在本向导全部运行完毕后，数据就会被转换成长型格式，可以发现转换后的数据和 anxiety. sav 基本上是相同的，同时结果窗口中也会给出相应操作的汇总表格用于查错。

4.2.4　数据转置

下面介绍转置(Transpose)过程，这实际上也是数据重构向导的第 3 个功能。转置过程用于对数据进行行列互换，即将记录转为变量，将变量转为记录后，重新显示在数据编辑窗口中，如图 4.16 所示。

	varname	v1	v2
1	A	1.00	2.00
2	B	3.00	4.00

(a)

	CASE_LBL	A	B
1	v1	1.00	3.00
2	v2	2.00	4.00

(b)

图 4.16　数据转置前后的数据文件

"转置"对话框也非常简单，如图 4.17 所示，左侧为候选变量列表框；右上方为"变量"列表框，用于选入需要转置的变量，一般应选入除名称变量外的所有其他变量，如果有变量未选入，则转置时会被自动丢弃；右下方为"名称变量"列表框，用于指定原数据文件中记录转置后变量名的字符变量，但不是必需的，此时系统会将新变量自动按 var001，var002，…的顺序命名。

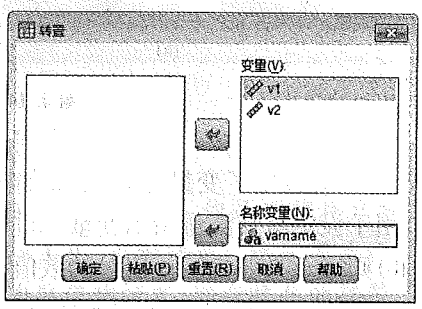


图 4.17　"转置"对话框

对统计分析的初学者而言，可能无法想象这个功能有什么用处。实际上，数据转置主要是用于编程进行矩

数据字典设定好，将来在数据录入完毕后使用复制文件属性向导即可套用字典。

（2）如果没有事先定义的数据字典格式，则可以在录入工作进行了一段时间以后先使用变量属性定义向导完成数据字典的设定工作，然后随着录入工作的进行扫描数据的情况，及时更新字典，最后在录入工作完毕后使用复制文件属性向导应用字典的最终版本。

（3）如果数据管理任务不太复杂，也可以直接在数据字典中录入数据，或者直接在变量视图中修改属性，或者直接在 SPSS 中录入/导入数据，然后利用设置未知测量级别向导来快速设定数据字典。但是在真正的大型数据管理项目中，单独建立和维护数据字典是非常关键的一环，建议有志于从事大型数据管理与分析工作的读者尽早养成这一良好的维护习惯。

4.4.2 定义变量属性

定义变量属性（Define Variable Property）指的是对于数据集中已存在的变量进一步定义其属性。具体说来，可以列出所选变量的所有取值；分辨没有值标签的值，并且提供自动给出值标签的功能；可以将另一个变量的属性复制给所选的变量，也可以将所选变量的属性复制给其他变量。从表面上看，该向导的绝大多数功能都可以在变量视图中实现，似乎有些多余，但对于复杂的数据管理项目而言，首先，它的可视化能力可以大大提高工作效率；其次，对初学者而言，使用该向导进行变量的设置也是非常好的选择。

这里仍以 CCSS 案例数据为例对该向导加以说明。选择“数据”→“定义变量属性”菜单项，则首先打开预定义对话框，要求选择希望进行设定的变量，可以选择多个，单击“继续”按钮后 SPSS 将会对选入的变量都进行扫描，随后进入向导的主界面，如图 4.22 所示。

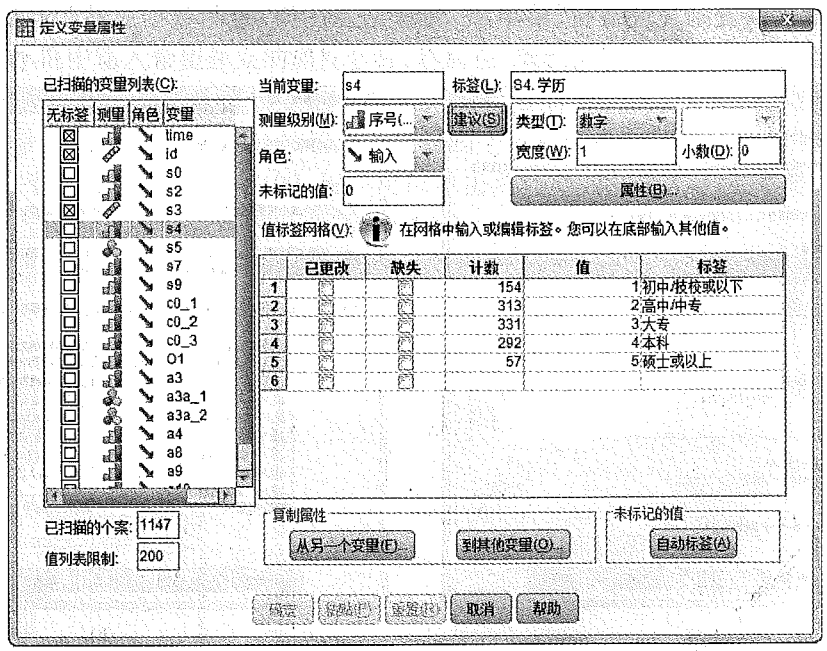


图 4.22 “定义变量属性”对话框

（1）左侧变量列表：其中会列出所有被选择/扫描的变量，具体又分为 4 列显示，分别表示该变量有无标签定义、测量尺度、角色和变量名。选中相应的变量名称，则右侧会显示出相应的设

定，并且允许用户更改。

（2）右上侧变量属性组：用于设定测量尺度、存储格式、变量名标签等，如果单击“建议”按钮，则系统会根据扫描到的数据给出建议的测量尺度；注意其中的“属性”按钮用于新建自定义属性，详见4.5.3小节的介绍。

（3）中部值标签网格：用于列出该变量所有取值的频数、当前值标签和缺失值设定等，双击后即可更改标签和缺失值设定（将当前数值设定为自定义缺失值）。

（4）下部“复制属性”框组：用于将另一个被扫描变量的属性复制给所选的变量，也可以将所选变量的属性复制给其他被扫描变量，该框组实际上已经在可视化离散过程中介绍过了。

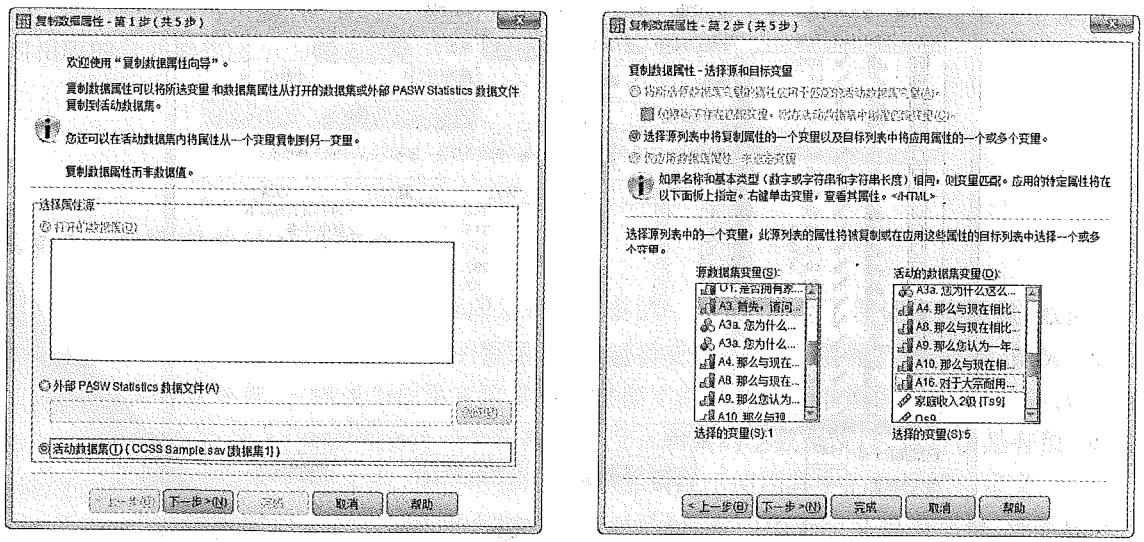
（5）“自动标签”按钮：用于自动生成值标签，实际上就是将所有的变量值均赋给空白值标签。

图4.21显示了S4的属性定义情况，由此可以看出在这一个界面中就可以完成对变量的所有属性定义，而且可以一次性定义多个变量，并且由系统帮助扫描出全部取值范围，这显然要比在变量视图中进行操作容易得多，可以大大方便数据字典的定义工作。

4.4.3　复制数据属性

复制数据属性（Copy Data Property）过程用于将定义好的数据字典直接应用到当前文件中，操作时不仅可以将一个外部数据文件的相关属性复制到当前数据文件中，还可以进行自定义，只选择某些变量，或者某些属性进行复制，这无疑大大提高了连续性项目对原有资源的利用程度。对于一些特殊的文件属性，如多选题变量集、普通变量集、权重变量设定等，使用该向导进行复制还会减少许多重复工作。

选择“数据”→“复制变量属性”菜单项，则首先会打开“复制数据属性”对话框，如图4.23所示。



(a)　　(b)

图4.23　“复制数据属性”对话框

（1）第1个对话框：在该对话框中可以选择希望复制的数据属性来源，可以是已打开的另一个数据文件，也可以是未打开的另一个数据文件，还可以是当前文件本身。这里选择了后者，然

后单击“下一步”按钮。

(2) 第 2 个对话框:第 2 个对话框用于设定希望复制的属性种类,有 3 种选择,分别为选择同名、同类型、同长度变量的属性进行复制(Apply properties from selected source file variables to matching working file),选择一个变量的属性进行复制(Apply properties from a single source variable to selected working file variable)和仅复制文件属性——多选题集定义、权重设定等(Apply dataset properties only—no variable selection)。在图 4.23 中选择的是第 2 项,具体设定是将变量 A3 的属性定义复制到 A4 ~ A16 这 5 个变量上去,单击“下一步”按钮。

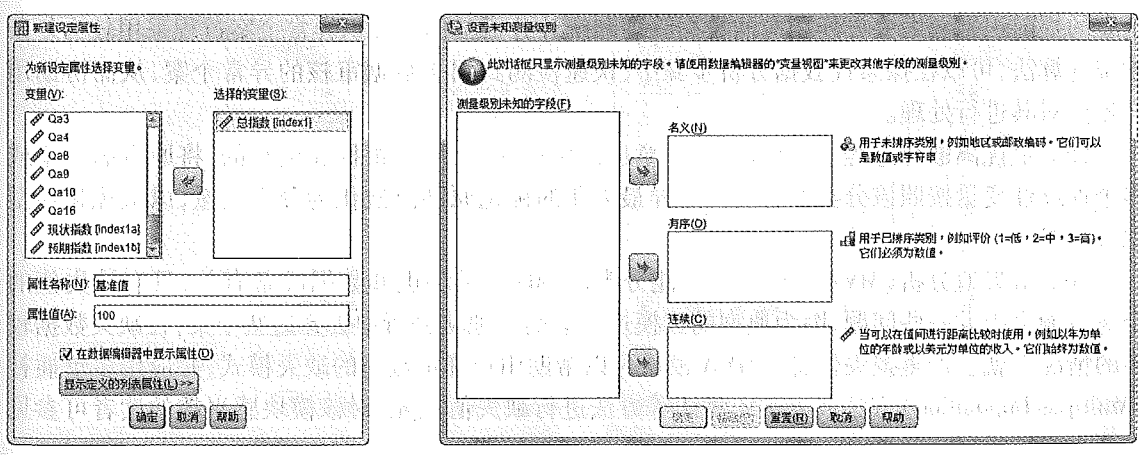
(3) 第 3 个对话框:该对话框要求详细指定希望复制的变量属性,共有 7 种,并且可以选择是替换原有属性,还是和原属性进行合并。

(4) 在第 3 个对话框出现时,用户其实就可以单击“完成”按钮结束向导了,第 4、5 个界面分别用于选择希望复制的文件属性,以及是否生成相应的 SPSS 程序。运行完毕后就会看到,A4 ~ A16 的变量属性全部更改成了与 A3 相同的设定。

4.4.4 新建自定义属性和设置未知测量属性

在默认情况下,SPSS 将为每个变量设定名称、类型等共 11 个属性,这在绝大多数情况下都是足够的,但是在一些大型数据管理项目中,可能需要用户自行设定一些特殊的变量属性,例如,可以创建识别调查问题类型的变量属性(单选、多选、开放)或存储计算变量使用的公式等。和标准变量属性一样,这些定制变量属性也将随数据文件一同保存。

在 SPSS 中建立自定义属性的操作非常简单,首先将数据窗口切换到变量视图,然后选择“数据”→“新建设定属性”(Create Custom Attribute)菜单项,打开的对话框如图 4.23 所示。在“属性名称”文本框中输入希望建立的属性名称,然后在“属性值”文本框中输入希望默认设定的属性值,并将希望进行默认属性值设定的变量选入右上方的“选择的变量”列表框中,例如,图 4.24 中选入的是 index1,单击“确定”按钮以后会看到变量视图最右侧新增一列“基准值”,其中 index1 对应的一行已经填充了数值 100,用户可以单击单元格右侧的省略号来增删该列的可能取值列表,对于空单元格也可以双击后直接输入信息。



(a) (b)

图 4.24 “新建设定属性”对话框和“设置未知测量级别”对话框

上面的内容非常容易理解，下面简要介绍各部分的功能。

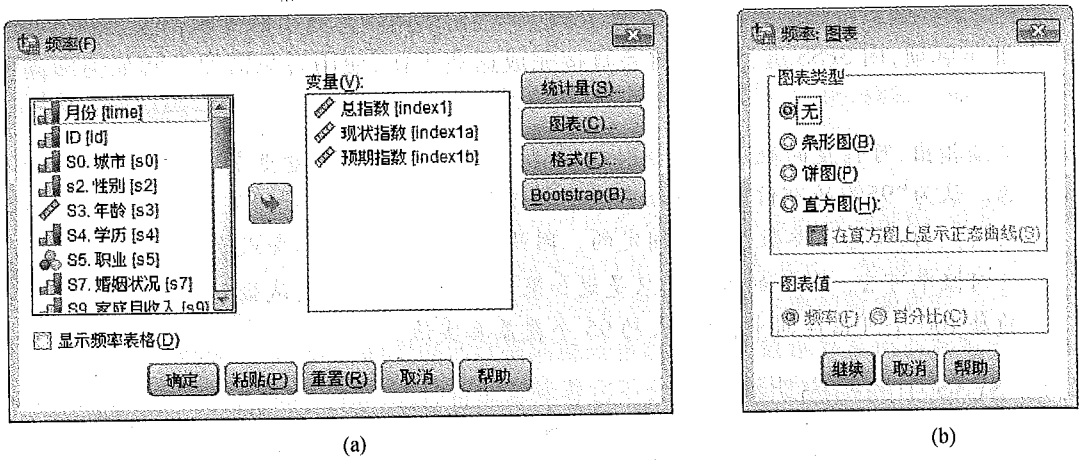


图 7.3　“频率”对话框和“频率：图表”子对话框

（1）主对话框：“变量”列表框用于选入需要进行描述的变量，如果选入多个，系统会依次对其进行分析。左下角的“显示频率表格”复选框用于输出频数表，默认选中。

（2）“统计量”按钮：单击后打开的对话框用于定义需要计算的描述统计量，包括集中趋势、离散趋势、分布特征和百分位数 4 组，比较特殊的是右侧的“值为组的中点”复选框，当输入的数据是分组频数数据，并且具体数值是组中值时，需要选中该复选框，这样 SPSS 在计算各种百分位数的时候会将数据按频数表对待，而不会认为同一组内的数据取值都是组中值的大小。

（3）“图表”按钮：单击后打开的对话框用于设定所做的统计图，可以绘制分类数据描述用的条图和饼图，也可以绘制连续变量描述用的直方图，相关的图形知识参见第 10 章。

（4）“格式”按钮：单击后打开的对话框用于定义输出频数表的格式，不过一般不用更改，使用默认设置即可。

（5）“Bootstrap”按钮：单击后打开的对话框可使用 Bootstrap 这种计算统计学方法进行任意总体参数的估计，详见 7.4 节的介绍。

2. 操作说明与结果解释

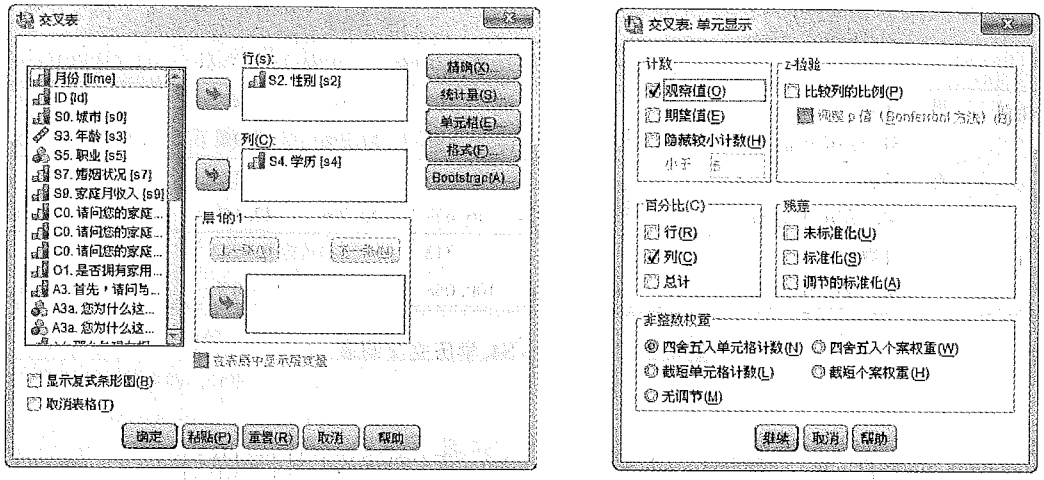
根据题目要求，本例操作如下。

（1）将 index1、index1a 和 index1b 选入“变量”列表框，取消左下方的“显示频率表格”复选框（因为本例不需要）。

（2）单击“统计量”按钮进入“统计量”子对话框，选中所需的常用统计量，并且在百分位数中设定输出 P2.5 和 P97.5，最终对话框界面应当如图 7.2 所示。

本例的输出结果如图 7.4 所示，可见总信心指数的均数和中位数非常接近，而根据百分位数法计算出的 95% 个体参考值范围为 46.86 ~ 132.78。大家如果有兴趣，可以利用均数和标准差计算出正态分布下的 95% 个体参考值范围是 54.74 ~ 137.05，显然和百分位数法的结果差异并不太大，这些信息都在暗示 index1 的分布可能是大致对称的。而用同样的方式可以发现现状指数的分布可能略呈偏态分布。

示,下面简要介绍各部分的功能。



(a) (b)

图 8.2 交叉表过程的对话框

(1) 主对话框:中部依次排列的"行"列表框、"列"列表框分别用于选择交叉表中的行、列变量。而下方的"层"框组则用于选入更多的分类变量作为层变量(详见第 9 章中关于表格结构的介绍),注意此处最多可进行多达 10 层的嵌套,同时行、列、层变量也是可以同时选择多个分类变量的。在左下角可以指定绘制复式条形图来呈现数据,而当交叉表太大的时候,也可以禁止表格输出。

(2)"精确"按钮:单击后打开的子对话框用于设定对行 * 列表是否进行确切概率的计算,以及具体的计算方法。本部分内容的介绍详见 12.5 节中关于蒙特卡罗方法的介绍。

(3)"统计量"按钮:单击后打开的子对话框中提供了一整套用于计算行/列变量关联性的统计指标和检验方法,详见卡方检验和相关分析两章中的介绍。

(4)"单元格"按钮:单击后打开的子对话框用于定义列联表单元格中需要显示的指标,这些指标被分为计数、百分比和残差三大类,实际上以前两类较常用。此外在现在的新版本中还提供了列于列之间进行对比的 Z 检验结果输出。

(5)"格式"按钮:单击后打开的对话框用于设定单元格的排序方式,使用价值不大。

(6)"Bootstrap"按钮:要求在相应的参数估计和假设检验中使用 Bootstrap 方法进行估计,该方法的原理和使用方式在第 7 章中已经介绍过了。

2. 操作说明与结果解释

根据分析目的,只需要分别将"性别"和"学历"选入"行"、"列"列表框中,然后在"单元格"子对话框中选择列百分比输出,即可得到所需的结果,如图 8.3 所示。

图 8.3 很清楚地给出了性别和学历的交叉分布情况,从中可以看出,随着学历的上升,男性所占的比例从初中/技校或以下的 48%,逐渐上升至硕士或以上的 63%。当然,由于这只是样本数据的描述情况,这究竟是因为抽样误差所致,还是总体中也的确存在此趋势,还需要通过假设

可能有的读者会对表格编辑有其他建议：为什么不将表格复制到 Word 里面进行编辑，Word 表格编辑起来不是更容易吗？是的，在 Word 里面操作的确更加容易，但并非结果表格最终都会放置在 Word 里面使用，粘贴到不同的软件里，就需要用不同的编辑方式来操作，因此学习 SPSS 的表格编辑操作更为稳妥。另一方面，许多编辑操作可以存储为表格模板以达到自动化出表的效果，这在实际工作中是非常有用的。

9.4.1 基本编辑操作

1. 两种不同的编辑窗口

在对结果表格进行编辑前，显然需要首先进入它的编辑模式。相应的操作非常简单，只需双击选中的表格，就会进入编辑状态。但由于 SPSS 的系统设置不同，可能是在新窗口中进入编辑模式，也可能是在结果浏览器中嵌套进入编辑模式。一般而言，对于较大的表格，单个窗口的编辑模式在操作上要更方便一些。如果希望能控制相应的编辑方式，除了可以在系统选项中加以设定外，还可以在选中相应表格后右击，在弹出的快捷菜单中选择“编辑内容”→“在阅读器中”或者“在单个窗口中”菜单项，前者使用嵌套模式，而后者将使用打开新窗口的方式进入表格编辑状态。

进入编辑状态后，在默认情况下窗口中会同时出现浮动的编辑工具栏和透视托盘，用于方便用户进行编辑操作。透视托盘在 9.1 节已经介绍过，用于控制和修改表格框架。通过工具栏则可以对选定单元格进行文字格式、对齐方式等的设定，单击其左侧的按钮可以切换透视托盘是否出现，单击右侧的按钮则可以根据表格数据生成统计图形，包括线图、饼图等，如图 9.19 所示。

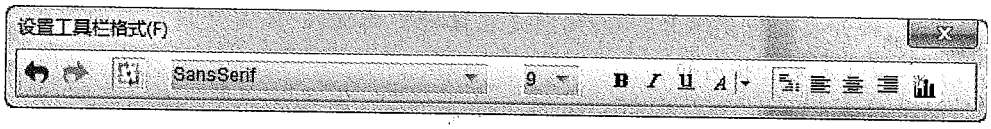


图 9.19 枢轴表的工具栏

工具栏自身是否出现是通过“视图”→“工具栏”菜单项切换的。

2. 表格元素的选择方式

在表格编辑中，单元格是基本的操作单位，包括表格标题和脚注均被看成特殊的单元格来处理。虽然根据所使用的表格模板设定不同，有的单元格间的分界线并未绘制出来，但它们在编辑操作中并不会被合并在一起，仍然是相互独立的编辑单位。

在对表格中的具体内容进行编辑操作时，显然应当首先将具体的元素选中，以使得系统得知相应操作是针对什么的。最常见的情形就是对单元格的选择，只需单击即可。不仅可以选中某个单元格，还可以选中其中的 1 行或 1 列，但首先要选中最上侧或左侧的标题单元格，然后选择“编辑”→“选择”菜单项，有 4 个选项：表格、表格主体、数据单元格、数据和标签单元格。在选中相应的单元格后，用户就可以对它们同时进行删除、复制、更改格式等操作，显然会方便得多。

3. 单元格内容的编辑

在题目 A3 生成的表格中，变量值标签还需要进行进一步修改。单击可以选中单元格，双击则进入单元格内数据的编辑状态，此时单元格内如果是数值，不仅会显示相应数据的全精度确切

1. 标题区和注解区

标题区和注解区这两个区域分别位于图形的最上方和最下方，位于图形最上方用于列出图形名称的就是标题区，如图 10.1 中写着“Count of …”的部分。标题中一般应注明图的编号，标题内容则简明扼要，用于说明资料的内容、地点、时间等。图 10.1 中最下方空白的区域即为注解区，主要用于添加对图形内容的简单说明，一般文字不宜过多，讲清楚即可。

需要指出的是，由于习惯不同，国内出版物中的统计图一般要求在图形的正下方给出标题，本书也是如此。在这种情况下，如果再添加注解，就会使图形显得不太对称，因此往往会将注解改为正文中的一段文字叙述。因此读者在实际绘图时，往往不会使用标题区和注解区的功能，而是自己另行添加。本书的实例也基本上不会用到这两个区域。

2. 坐标轴

坐标轴、图形本身（绘图区）在内的区域一般被统称为数据区，是统计图的主要部分，这里将分开讲述。坐标轴用于表示相应变量的取值情况，由于二维统计图最为常用，相应的两个坐标轴往往被直接称为横轴和纵轴。实际上应当按照所表示的数据类型将坐标轴分为连续轴和分类轴两大类，如图 10.1 中所示的横轴就是分类轴，其数轴刻度间无大小之分，仅代表不同的类别。其纵轴则为连续轴，刻度严格而准确地表示了数量上的差异。连续轴和分类轴的编辑功能相差极大，但与其位于横轴还是纵轴则完全无关。

坐标轴一般都应注有标目，用于说明其表示的具体含义。对于连续轴而言，往往还需要注明单位，如年份、克、% 等。连续轴的刻度设定应该是等距的，而且在一般情况下为算术等距，但必要时也可以是几何等距，以满足特殊的分析需求，如图 10.2 所示。纵横尺度一般从 0 开始（对数线图、点图例外），以免曲解统计图所表示的指标关系。

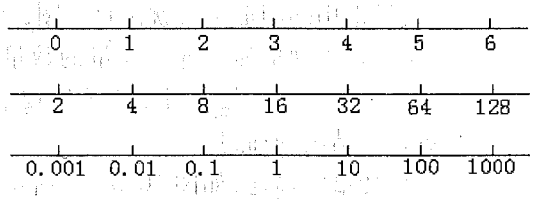


图 10.2 算术尺度、几何尺度和对数尺度的连续轴

3. 绘图区

绘图区指的是被坐标轴包围，直接使用图形元素来对数据进行呈现的区域，在 SPSS 中也被称为内框区，以和表示整个图形范围的外框区相对应。绘图区中主要有表示变量数值情况的直条、区块、点、线等图形元素，使用者在阅读图形时需要首先注意相应的各坐标轴的具体含义，以明确各图形元素的坐标究竟表示的是数量大小，还是类别。如图 10.1 中所示的不同直条（横轴为分类轴）表示的是汽车的不同产地，而直条的高低（纵轴为连续轴）则表示了具体指标（这里为 MPG、ACCEL）的算术均数。

除了基本的图形元素外，绘图区中还可能出现各种文字注解、辅助坐标线等用于方便图形阅读的元素。

4. 图例区

图例区位于整个图形的右侧，当在图形中需要使用不同的颜色、线形等将图形元素分组以表

该对话框被关闭，则可以选择“编辑”→“属性”菜单项将其重新打开。

2. 图形编辑的基本操作要点

这里首先列出图形编辑的基本操作要点如下。

(1) 选择图形元素：统计图中的各种图形元素，如散点、直条、数轴等，都会按照其统计特征进行编组，在选择这些元素时，基本规律是第一次单击会选中图中的所有同类/组元素；在原位置上两次单击，则会变为只选中该图形元素本身(在使用图例时，两次单击会选中同组元素，3 次单击才会选中元素本身)。如果希望选择不同的多个图形元素，则按住 Ctrl 键分别选择即可。对于所选中的一个或者一组图形元素，用户可同时对其进行相同的格式编辑操作，如颜色、填充样式，甚至于单独标出具体的数值、ID 号等。

(2) 进行文本编辑：首先用鼠标单击选中文本，对于可编辑的文本元素，再次单击则进入编辑状态，可对其自由进行内容、格式、字体等的编辑。

(3) 移动图形元素或改变其大小：同样需要先用鼠标选中相应元素，然后视选中框四周是否出现控制柄来确定该元素能否被移动或改变大小，对于无控制柄的选中框，相应元素是无法移动的，否则就可以移动或改变大小，详见下面介绍。



文本的编辑设定一般分为 3 种情况：对于数轴刻度等移动/更改可能导致图形误读的文本元素，既不能移动，也不能编辑；对于图例文字等，则可以编辑，但不能在图例内部进行相对位置的移动；其他不太重要的文本元素则既可以随意移动，也可以进行内容编辑。

图形元素的位置设定也大致分为 3 种情况：数轴刻度等重要图形元素的位置完全固定，以保证基本的图形特征；图形中的坐标参考线等的位置为半固定，不能使用鼠标随意拖动，而只能在选项卡中输入坐标进行精确定位，以保证位置的准确；而对于一般性的文字注解、标签等，用户可以在图形中进行随意拖动。

3. 更改图形长宽比例

默认图形是按照 1:2 的长宽比绘制的，相对而言有点窄。如果希望更改为更宽的 3:4 等比例，则在“属性”子对话框的“图表大小”选项卡中取消选中“保持宽高比”复选框，然后在上方的“高度”、“宽度”数值框中自由输入希望设定的图形大小，再单击“应用”按钮即可，本例更改为 12:16。

4. 图例元素的位置移动和改变大小

除长宽比不合适外，默认绘制的直方图由于右侧完全被统计量图例占据，导致
起来更窄，完全可以将该图例拖动到更合适的位置上去：用鼠标将整个图例元素
出现如图 10.10(a)所示的带 8 个控制柄的方框，将光标移动到框线上，则光标
形，此时按下鼠标左键即可随意移动所选中元素的位置。如果将光标移
光标变为双向箭头形，此时按下左键可以更改元素的大小，如图 10.10
置或改变大小时相应区域内的文本大小和格式设置不会改变，只
动”。而其中的图形元素则会自动调整大小和形状，如改变直
佳的显示效果。

线以突出
选择“添加 X 轴
即 80 的位置，因
75
元素可被删除，则在菜单上
5
间。但其功能和图形需求是完全

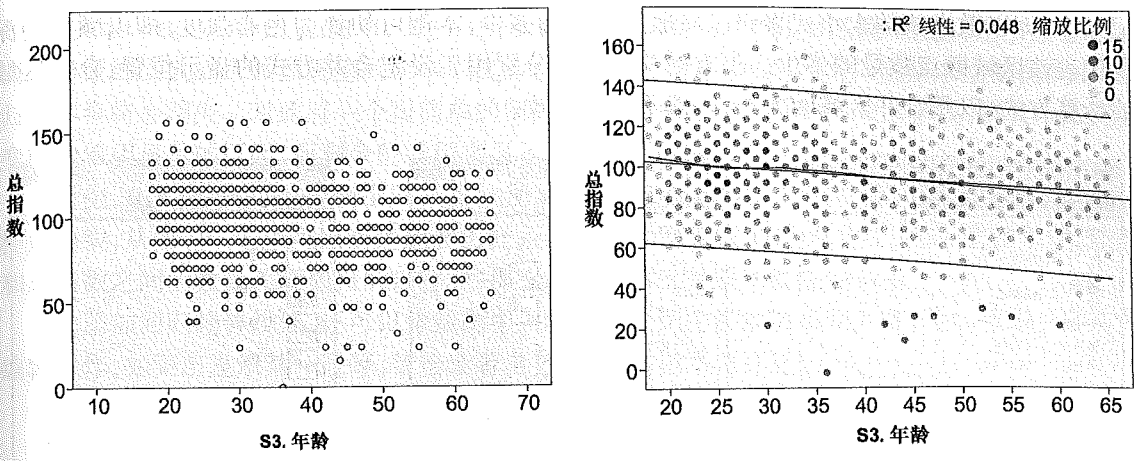


图 10.31 默认绘制生成及编辑完毕后的简单散点图

如果进一步进行两变量间的回归或相关分析，上述信息将对分析工作起到重要的方向指导作用。

10.7.2 散点图的编辑

散点图中的图形元素以散点为主，因此前面讲述过的各种散点编辑功能，如更改散点样式、大小、只显示某些散点的标签等均可以加以应用。除此以外，散点图中还有一些独特的编辑功能，如更改散点密度的显示方式、加入回归趋势线等，这里将结合 10.7.1 小节的实例来加以讲解。

1. 用套索模式选择离群散点

在图 10.31 所示的散点图中可以看到下方有一些偏离稍远的点，可以将其显示方式编辑得更为醒目一些。但由于数量太多，如果采取按住 Ctrl 键依次选中的方式显然太麻烦，这里可以采用套索模式来实现。选择“编辑”→“套索选择标记”菜单项，或者直接单击工具栏上的按钮，可以看到鼠标变成套索形状，此时在散点图中按下鼠标左键，将希望选中的散点圈入一个闭合曲线环中，松开左键后就会看到这些散点已被同时选中，下面就可以对这些散点同时进行各种编辑操作了。

较新的版本

套索模式是19版新增的，老版本无此选项。

2. 改变过密散点图的显示方式

CCSS 案例数据的样本量虽然上千，但由于年龄只能取整数，信心值也并非任意取值，导致所绘制的散点图中有大量散点是重叠显示的，各部分的疏密无法分清，这会严重影响对散点图趋势的观察。

事实上，在大样本数据，或者变量取值范围有限的情况下，散点重叠的情形非常常见，对此问题可以采用散点合并的显示技巧加以解决。在编辑状态下选择“选项”→“块元素”菜单项，或者在快捷工具栏上直接单击按钮，就可以将散点转化为合并显示。同时在“元素”对话框中会增加一个“分箱”选项卡（如图 10.32(a)所示），用于设定具体的合并选项。默认的显示方式为标

属性

散点图矩阵的对角线处实际上会显示该变量的直方图，只是在默认情况下是隐藏的，如果希望显示，则选择“选项”→“显示沿对角线绘制的图表”菜单项即可。

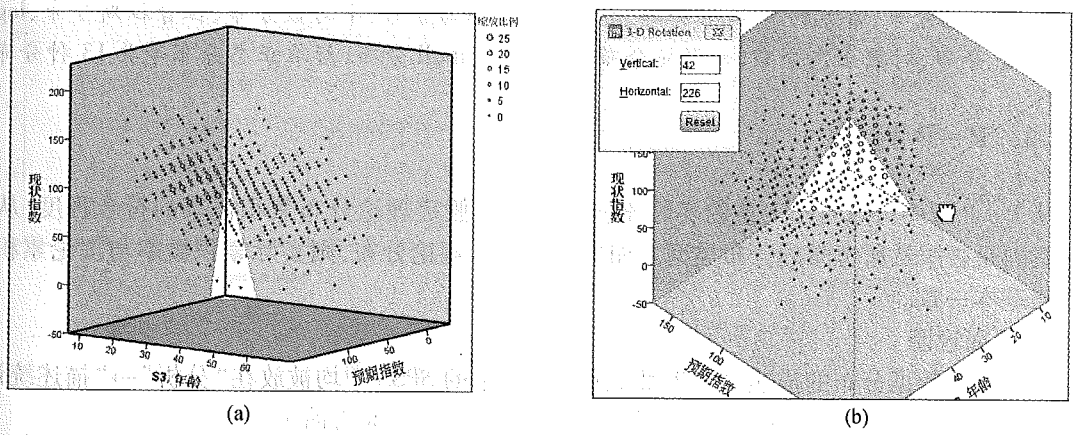
从散点图矩阵中可见，年龄与现状指数和预期指数均呈负相关关系，年龄越大，指数值越低，但似乎年龄与现状指数之间存在一定的曲线关联，在后续分析中要加以注意。

10.7.5 三维散点图

在散点图矩阵中虽然可以同时观察多个变量间的联系，但是两两进行平面散点图的观察的，有可能漏掉一些重要的信息。三维散点图就是在由 3 个变量确定的三维空间中研究变量之间的关系，由于同时考虑了 3 个变量，常常可以发现在两维图形中发现不了的信息。

1. 三维散点图的绘制

仍以上面的问题为例，如果希望直接做出 S3、index1a、index1b 的三维散点图，则只要在对话框中将它们依次定义为 X、Y、Z 轴即可，所绘制出的三维散点图如图 10.34 所示。该图形将 3 个变量间的关系在同一个坐标空间中立体地表现了出来，使用它可以更加清晰和直观地对应/自变量间的关系进行观察，发现在二维空间中可能无法看到的信息，如曲线关系、异常值等。但是，由于实际上只能在二维平面上观察三维散点图，所以在观察时必须要结合旋转功能，这将在下面加以讲述。



(a) (b)

图 10.34 三维散点图及其自由旋转操作示例

2. 三维散点图的旋转观察

对三维散点图的旋转进行观察需要进入编辑状态，然后选择“编辑”→“3D-Rotation”菜单项，或直接单击工具栏上的按钮，就会看到出现一个 3D 旋转托盘，如图 10.34 所示，用它进行旋转，可马上看到旋转的效果。具体操作时既可以在对话框中更改纵横坐标，也可以按下鼠标左键对三维图形做各个方向的直接拖动，即压住左键，然后将鼠标向需要的方向移动，图形就会向相应方向转动，直至松开鼠标左键为止。

3. 三维散点图的缩放

在默认情况下，三维散点图主体只占据了图形的一部分区间，这是由于默认对散点图的观察距离较远所致。如果希望散点图占据主要显示区间，则可以选中散点图主体，然后在“3-D 旋转”选项卡中将距离改得较近一些即可，默认为 70，一般改为 30 即可使图形占据大部分显示面积。

2. 对性别和职业的影响进行分析

下面对性别和职业的影响进行分析,这里仍然首先采用均数过程加以描述,因操作与前面基本相同,此处省略相应的说明。

图11.5暗示不同性别的受访者其信心指数似乎没有太大差异,该结果应当符合整个项目的设计,毕竟所询问的是整个家庭的情况,从逻辑上讲不同性别的受访者其感受是不应当有太大差异的。

均值

S2. 性别	总指数	现状指数	预期指数
男	95.9719	99.8869	93.8160
女	95.7956	98.3931	94.3643
总计	95.8935	99.2227	94.0598

图11.5　总指数 现状指数 预期指数 * S2. 性别

图11.6显示出在本科学历之前,随着学历的上升,受访者的总信心指数是上升的,随后在硕士级别开始下降。分指标的描述结果表明,实际上现状指数基本上是随着学历的上升而上升的,但预期指数则在硕士及以上级别出现明显下跌。

均值

S4. 学历	总指数	现状指数	预期指数
初中/技校或以下	93.6773	94.6100	93.1620
高中/中专	94.6264	97.6692	92.9502
大专	96.6305	101.6669	93.8576
本科	97.6066	100.1710	96.1936
硕士或以上	95.7835	101.1641	92.8213
总计	95.8935	99.2227	94.0598

图11.6　总指数 现状指数 预期指数　* S4. 学历

可以进一步考虑将性别和学历进行联合描述,这里考虑使用多重线图来完成此任务,操作步骤如下。

(1) 选择“图形”→“图表构建程序”菜单项,打开“图表构建程序”对话框。

(2) 选择“线”图组,将多重线图拖入画布中。

(3) 将“学历”拖入横轴框中。

(4) 同时选中“总指数”、“现状指数”、“预期指数”,将其拖入纵轴框中。

(5) 选择“组/点ID”选项卡,选中列嵌板变量。

(6) 将“性别”拖入画布中新增的“列嵌板变量”框中。

(7) 单击“确定”按钮。

从图11.7中可以发现,虽然总体而言不同性别受访者的信心水平没有明显差异,但是当性别和学历交叉之后,性别×学历对信心指数的反映并不相同,大致有如下趋势。

12.3 二项分布检验

对于两分类变量而言，二项分布是最常见的分布类型，本节就来介绍一下二项分布检验方法。

12.3.1 二项分布检验的原理

二项分布检验(Binomial Test)是对二分类变量的拟合优度检验，用于考察每个类别中观察值的频数与特定二项分布下的预期频数间是否存在统计学差异。二项分布检验的原理实际上和K-S检验的原理相同，只是这里使用的是二分变量，是一个离散分布的检验情况。

在第8章中已经介绍了二项分布的基本知识，对于一个服从二项分布的随机变量而言，在 n 次试验中结局 A 出现的次数 X 的概率分布为

$$P(X=k)=\binom{n}{k}\pi^{k}(1-\pi)^{n-k} \quad k=0,1,\cdots,n$$

使用上述公式就可以算出基于无效假设时各发生次数的出现概率，利用小概率反证法，按照和K-S检验中类似的逻辑给出相应的检验结论。

12.3.2 案例：考察抽样数据的性别分布是否平衡

例12.2 在人群中性别比例基本上是1∶1，考察CCSS样本中的数据是否仍然符合此规律。

本例所对应的检验假设如下。

H_0：男性(或女性)比例 $\pi=0.5$，样本所对应的总体男女比例一致。

H_1：$\pi\neq0.5$，男女性比例不一致。

注意此处仍然只使用2007年4月的样本数据进行分析，具体的分析操作也仍然使用“单样本非参数检验”对话框来完成。

(1) 选择“数据”→“选择个案”菜单项。

(2) 在打开的对话框中选择框组：如果条件满足。

(3) “如果”子对话框：time=200704。

(4) 选择“分析”→“非参数检验”→“单样本”菜单项，打开“单样本非参数检验”对话框。

(5) “目标”选项卡：自动比较观察数据和假想数据。

(6) “字段”选项卡：使用定制字段分配，将S2性别选入“检验字段”列表框中。

(7) “设置”选项卡：选择“自定义检验”选项组中的第一项“二项式检验”，相应选项中“假设比例”已经是所需的0.5了，因此不需要更改，如图12.7所示。

(8) 确定。

最终的模型输出如图12.8所示，右侧辅助视图中同时给出了165例的男性样本频数，以及标化统计量1.674的数值，最终基于样本和无效假设推导出的 P 值为0.094，因此当无效假设成立时，100次中平均有9次可以得到偏离理论值150例和现有样本一样远甚或更远(包括大于等于165例，或者小于等于135例两种情况)的样本，按照默认的0.05水准，这并非小概率事件，因

在零假设成立时，F 值应该服从自由度为 $k-1$，$N-k$ 的中心 F 分布（Central F Distribution）。自由度为 1，5 的 F 分布如图 14.1 所示。而若检验统计量落在相应检验水准所确定的拒绝域内（即 F 值大于或等于相应自由度下的检验界值），意味着在一次抽样研究中在假设总体内得到了小概率事件，则有理由拒绝 H_0，其风险为相应 F 值所对应的 P 值。

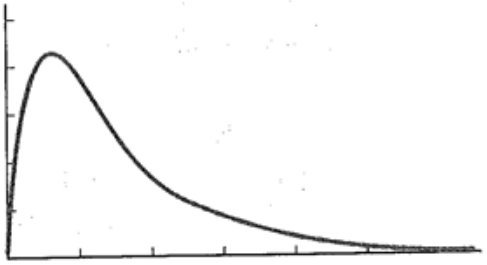


图 14.1 自由度为 1，5 的 F 分布

在方差分析中常常将所计算出来的一些指标列成一张表格，称为方差分析表（Analysis of Variance Table），如表 14.1 所示。

表 14.1 方差分析表

变异来源	离差平方和	自由度	均方	F	P
组间变异	SS_B	$k-1$	MS_B	MS_B/MS_W	$P=\{F_{k-1,k(ni-1)}\geqslant F\}$
组内变异	SS_W	$k(n_i-1)$	MS_W		
总变异	SS_T	$N-1$	MS_T		

后面还会讲到，SPSS 的输出结果中将包含一张与此几乎完全相同的表格。

如果假设检验拒绝了 H_0，可以得出多个样本不是来自同一总体的结论。但是到底这些样本来自于几个不同的总体，这次假设检验还不能回答这个问题，而需要进一步进行单因素不同水平间的多重比较（Multiple - Comparison），详见后面讲解。

14.1.3 单因素方差分析的应用条件

1. 基本的应用条件

统计学中也许有成千上万的方法，但没有哪种方法是通用的，自然这里的方差分析也不例外。一般而言，要应用方差分析，数据应当满足以下几个条件，或者说以下的假设应当成立。

（1）观察对象是来自于所研究因素的各个水平之下的独立随机抽样（Independence）。

（2）每个水平下的应变量应当服从正态分布（Normality）。

（3）各水平下的总体具有相同的方差（Homoscedasticity）。

其实，与 t 检验的应用条件大同小异，概括起来就是独立性、正态性和方差齐性。

2. 应用条件的检查与变量变换

以上适用条件可以使用统计描述进行观察，或者绘制相应的统计图形，当然也可以使用相应的检验方法。正态性检验的方法不再介绍，这里简要介绍 3 个或 3 个以上样本的方差齐性检验方法。

（1）Bartlett 法：其基本思想是比较各组方差的加权算术均数与几何均数，若两者差异过大，可以认为各组间的方差不齐。当各组样本含量均大于 5 时，其检验统计量近似服从自由度为

图 14.13 所示的方差分析表中分别对均数变化趋势是否服从线性、二次多项式的方程进行了检验，可见线性项、二次项检验所对应的 P 值均小于 0.05，因此，从统计检验的角度来看，总指数和月份之间的关系不能用线性关系来描述。但这里需要说明的是，虽然此处的检验结果拒绝了相应的趋势假设。但是多项式形式的判定并不是完全靠 P 值来决定的，许多时候还要依靠图形、专业知识以及经验来给出结论。

总指数

			平方和	df	均方	F	显著性
组间	(组合)		9426.352	2	4713.176	10.884	.000
	线性项	加权的	7309.948	1	7309.948	16.881	.000
		偏差	2116.404	1	2116.404	4.887	.027
	二次项	加权的	2116.407	1	2116.407	4.887	.027
组内			391895.938	905	433.034		
总数			401322.290	907			

图 14.13 ANOVA

表 14.9 中的“加权的”输出指的是对样本是否符合当前模型假设进行的检验，而“偏差”则指的是当前模型和含有最高次项的模型相比是否有区别，换言之，就是当前模型是否还需要继续增加高次项。

14.6 本章小结

(1) 单因素方差分析所针对的是多组均数间的比较。它的基本思想是变异分解，即将总变异分解为组间变异和组内变异，再利用 F 分布做出有关的统计推断。

(2) 单因素方差分析要求资料满足正态性、独立性和方差齐性的要求。

(3) 方差分析拒绝 H_0 只能说明各组之间存在差异，但并不足以说明各组之间的关系。利用多重比较可以初步判断各组间的关系。

(4) 多重比较可以分为事前计划好的比较和事后比较。前者往往借助于 Contrast，而后者有很多种不同的方法，这些方法的核心问题都是如何控制总的一类错误的大小。

(5) 在分组变量包含次序信息时，如果方差分析给出了各组间差异有统计学意义的结论，并且 Means-Plot 提示各组均数的某种趋势时，可以利用趋势分析探讨观察值与分组变量取值间的数量依存关系。

思考与练习

1. 一家汽车厂设计出 3 种新型号的手刹，现欲比较它们与传统手刹的寿命。分别在传统手刹，型号 I、II 和型号 III 中随机选取了 5 只样品，在相同的试验条件下，测量其使用寿命(单位：月)，结果如下。

和图 15.4 所示的新对话框相比，旧对话框有两个优势：可以计算确切概率值，以及可以同时进行多对变量的检验。

单击“确定”按钮，即得到分析结果，如图 15.5 所示。

		N	秩均值	秩和
清华录取平均分 - 北大录取平均分	负秩	17[a]	16.47	280.00
	正秩	14[b]	15.43	216.00
	结	0[c]		
	总数	31		

a. 清华录取平均分 < 北大录取平均分。
b. 清华录取平均分 > 北大录取平均分。
c. 清华录取平均分 = 北大录取平均分。

图 15.5　秩

图 15.5 中列出了对应于两所高校录取平均分数差值为正、负、相等 3 种情况时的秩频数、秩均值及秩和。

图 15.6 给出的分析结论与前面的模型输出完全一致，不再重复解释。

	清华录取平均分 - 北大录取平均分
Z	-.627[a]
渐近显著性(双侧)	.531

a. 基于正秩。
b. Wilcoxon 带符号秩检验。

图 15.6　检验统计量[b]

1. 符号检验的分析结果

如果在例 15.1 中使用符号检验，则分析结果如图 15.7 所示，可见近似 P 值等于 0.719，虽然在本例中结论相同，但是，两者的近似 P 值有比较大的差别。这主要是由于符号检验对信息的利用程度不如符号秩检验，检验效能不够充分所致。

在绝大多数情况下，对信息利用程度较差的方法所得到的 P 值会更大，但也的确会存在相反的情形。

	清华录取平均分 - 北大录取平均分
Z	-.359
渐近显著性(双侧)	.719

a. 符号检验。

图 15.7　检验统计量[a](符号检验)

另两种方法因在本例中不满足使用条件，此处不再给出结果示例。

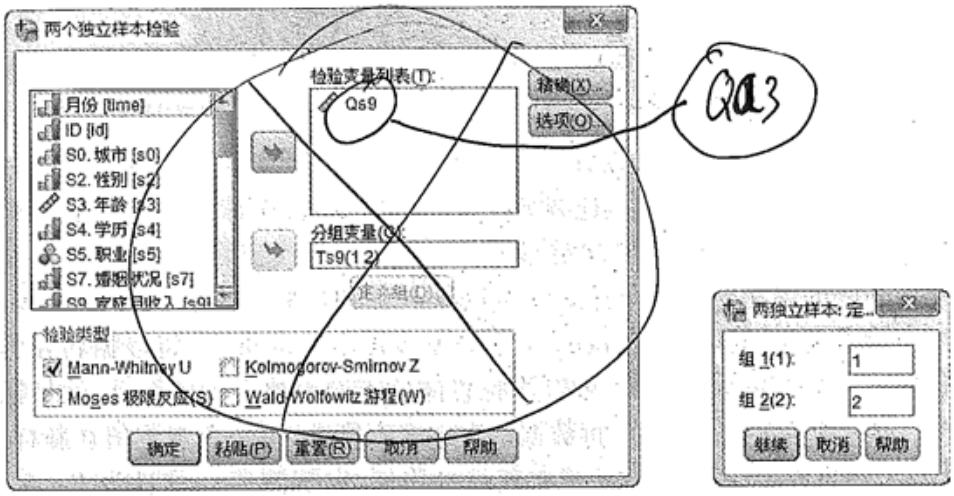


图 15.11 “两个独立样本检验”对话框

相应的操作如下。

(1) 选择“分析”→“非参数检验”→“旧对话框”→“两个独立样本”菜单项,打开“两个独立样本检验”对话框,进行如下设置。

“检验变量列表”框:Qa3。

“分组变量”列表框:Ts9。

(2) 单击“定义组”按钮,在打开的对话框中设置组 1 为 1,组 2 为 2,单击“继续”按钮。

(3) 单击“确定”按钮。

相应的分析结果如图 15.12、图 15.13 所示。

	家庭收入 2 级	N	秩均值	秩和
Qs9	Below 48,000	110	55.50	6105.00
	Over 48,000	145	183.00	26535.00
	总数	255		

图 15.12 秩

	Qs9
Mann-Whitney U	.000
Wilcoxon W	6105.000
Z	-13.738
渐近显著性(双侧)	.000

a. 分组变量:家庭收入 2 级。

图 15.13 检验统计量[a]

15.4 多个独立样本的非参数检验

多样本问题主要涉及如何检验几种不同的方法、决策或处理所产生的结果是否一样;商业活动中采取不同的决策方案风险的大小是否有区别;不同的销售差异,其节生活中不同的消费者对不同的产品偏好是否有显著差异;不同的运动方式或饮食对 12 月和 2008 年否相同。在第 14 章中进行多组均数的比较时,利用了方差分析来推断 3 个或检验结果可以看出,值的相等性。但是该过程需要若干条件,如要求间隔或比例数据,所有总差异。从实际含义角度总体的方差均相等。但是有时候所采集的数据常常不能满足这些条件,入谷底,其后虽然总信心指不满足都会使人陷入尴尬之中。当不满足这些条件时,F 检验就受

际上本例中数据的一致性并不是很强。特别是有一个地址两人竟给出了完全对立的评价。

一致性检验在医学研究中用得很多。如在一种简单易行的诊断方法是否可替代另一种结果可靠但操作繁杂的诊断方法的研究中，就会用到一致性检验。另外，在数据分析中，比较两种预测方法预测结果的一致性时也可能会用到 Kappa 检验。

16.5.2　配对卡方检验

通过 Kappa 检验已经回答了两种测量间究竟有无关联的问题，但是通过对列联表的观察会发现，两位顾问的评价似乎不太一样，这种问题又如何来加以分析？McNemar 配对卡方检验就是经典的配对检验，专门用于解决此类问题。在统计量子对话框的左下角就是 McNemar 复选框，例如上例，选择后相应的结果输出如图 16.13 所示。

	值	df	渐进 Sig.（双侧）
McNemar-Bowker 检验	8.000	3	.046
有效案例中的 N	20		

图 16.13　配对卡方检验

此处的原假设为：两顾问的评价结果无差别，显然，P 值小于 0.05，因此拒绝了该假设，认为应当是有差别的，从样本数据看，应当是第一个顾问倾向于评价得更高。



对于 4 格表数据，配对卡方会直接使用等价于二项分布检验的 McNemar 检验给出确切 P 值。但是对于较大的表格，则只能使用近似的 McNemar-Bowker 检验给出近似 P 值。

现在，Kappa 检验显示两者的评价存在一致性，而配对卡方检验则显示两者的结果是有差别的。实际上，这两个结论并不矛盾，参考前面对 Kappa 值的评价方式就可以理解。另外，这两者在信息的利用上也有差异。Kappa 检验会利用列联表中的全部信息，而 McNemar 检验只会利用非主对角线单元格上的信息，即它只关心两者不一致的评价情况，用于比较两个评价者间存在怎样的倾向。



在应用中，对于一致性较好，即绝大多数数据都在主对角线上的列联表，McNemar 检验可能会失去实用价值。例如对 1 万个案例进行一致性评价，9 995 个都是完全一致的，在主对角线上，另有 5 个分布在左下侧的三角区，显然，此时一致性相当好。但如果使用 McNemar 检验，由于它并不考虑主对角线上的数据，只会利用上、下三角区的信息，此时反而会得出两种评价有差异的结论。而正确应用这些方法的关键点就在于弄清楚自己希望考察的究竟是一致性还是差异性。

16.6　分层卡方检验

在例 16.2 中，经卡方检验发现家庭收入级别的确会影响家庭轿车的拥有情况，随后又进一步计算出了两者的关联强度指标 *OR* 和 *RR*。但在进一步考虑之后，研究者会发现还存在如下问题。

（1）不同城市的轿车拥有情况是存在差异的，那么收入级别对其的影响在不同城市间是否也存在差异？比如说在有的城市影响大些，而在其他城市的影响小些？

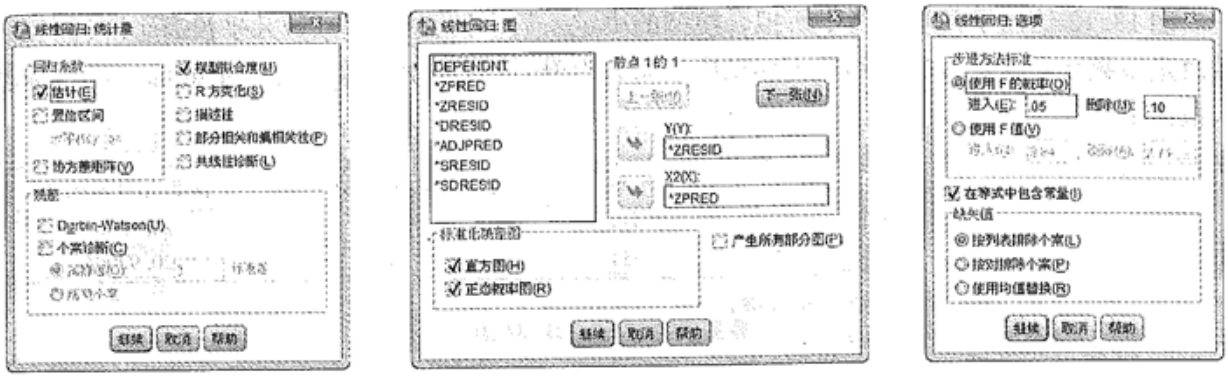
2. 在周六晚节目单修订前后，分别进行了收视率的调查。在节目被修改前，300个家庭收视纪录为：ABC76个家庭，CBS89个家庭，NBC83个家庭，独立电台52个家庭。节目被修改后，300个家庭所组成的样本产生下列电视收视数据：ABC95个家庭，CBS70个家庭，NBC89个家庭，独立电台46个家庭。取显著性水平 $\alpha=0.05$，检验电视收视率是否已经发生了变化。用软件SPSS进行分析，并解释各表的含义（将本题与第1题进行比较）。

3. 3名推销员3个月内的销售数量报告如题表所示。取显著性水平 $\alpha=0.05$，检验推销员与产品类型的独立性。

题　表

推销员	产品		
	A	B	C
Michael	14	12	4
David	21	16	8
Alice	15	5	10

4. 一家生产性公司从3家供应商处购买某零件，但该零件经常出现次品。在记录的435件零件质量数据中，100件来自A公司，其中90件质量等级为良好，3件有小缺陷，7件有大缺陷；195件来自B公司，其中170件质量等级为良好，18件有小缺陷，7件有大缺陷；150件来自C公司，其中135件质量等级为良好，6件有小缺陷，9件有大缺陷。取显著性水平 $\alpha=0.05$，检验供应商与零件质量的独立性。分析结果能给采购部门提供什么信息？



(a) (b) (c)

图 18.10 线性回归过程的“统计量”、“图”和“选项”子对话框

2. 自动线性建模过程

该过程是 SPSS 近几个版本向自动化、智能化操作平台转化的成果之一，利用该过程，用户可以采用几乎完全自动的方式进行自变量的预变换、筛选、模型优化、检验等工作。自变量也可以是连续、有序、无序等任何一种测量尺度，系统会自动选择相应的转换方式/算法来加以分析。

由于自动准备数据过程的自动化程度较高，中间又会涉及一些比较复杂的算法，因此笔者并不建议不了解回归模型细节的初学者使用，本书将不对其进行深入介绍，对该过程感兴趣的读者可参考《SPSS 统计分析高级教程》。

18.3.5 案例：建立自变量包括年龄、性别、家庭收入的信心指数回归方程

例 18.1 建立候选自变量包括年龄、性别、家庭收入的总信心指数回归方程。

本例需要建立包括 3 个候选自变量的回归方程，由于事前不能确定这些自变量是否均具有统计学意义，因此出于简化模型的思路，可以考虑用向后法进行变量筛选。

1. 操作说明与结果解释

(1) 选择“分析”→“回归”→“线性”菜单项，打开“线性回归”对话框。

(2) 将“总指数”选入“因变量”列表框中，将“年龄”、“性别”、“Qs9”选入“自变量”列表框中。

(3) 在“方法”下拉列表框选择“向后法”选项。

(4) 单击“确定”按钮。

由于进行了变量筛选，最终输出的表格较多，依次解释如下。

图 18.11 给出了 SPSS 在回归过程中每个步骤所进行的操作，第 1 步是采用进入法选入了全部 3 个候选自变量，然后在第 2 步剔除了性别，原因是其检验概率大于 0.1 的剔除标准。

模型	输入的变量	移去的变量	方法
1	Qs9，S2. 性别，S3. 年龄		输入
2		S2. 性别	向后(准则：F-to-remove >=. 100 的概率)。

a. 已输入所有请求的变量。

b. 因变量：总指数。

图 18.11 输入/移去的变量[b]

存在着高度的离散性,很可能不服从正态分布。实际上,这恐怕是所有金融类模型都可能遇到的问题。这里同样不考虑进行事前处理,而是建模之后再进行考虑。

最后,从图19.37所示的数据可以得出一个关于该行业的结论:低市场集中度和低利润率并存是我国证券业的一大特点。

2. 数据变换

下面考虑对建模所需的变量进行计算,根据上面的讨论,相应的数值计算程序如下:

```
COMP 经营效率 = 营业收入/总资产.
COMP ACR = 总资产/所有者权益.
EXEC.
```

上述程序运行之后,数据集中就会增加经营效率和ACR两个新变量,注意这两个新变量中也存在缺失值,这是因为计算用原始变量也存在缺失值。

19.4.4　数据建模



比较熟悉统计模型的读者可能会想到,这应当是一个比较典型的重复测量数据结构:20家公司在6个年度重复给出所测量的数据,构成了一个完美的重复测量结构,因此在分析的时候就应当考虑这种特殊的数据结构,采用重复测量的一般线性模型,最好能够采用重复测量的GEE模型或者混合效应模型来进行建模估计——这类想法完全可以理解,但模型是用来解决问题的,本案例分析目的如此简单,有必要建立这么复杂的分析模型吗?

下面考虑基于上述数据对前面构建的模型公式进行估计,采用回归过程来分析,注意此处不应当采用变量筛选,而是使用进入法将所有变量纳入模型,相应的结果如图19.45所示。

图19.45显示整个模型的决定系数为39%,也就是说对样本的因变量变异有一定的解释能力,但是否具有统计学意义尚需进行检验。

模型	R	R方	调整R方	标准估计的误差
1	.624[a]	.390	.333	.16951

a. 预测变量:(常量),证券市场增长率,ACR,市场份额,经营效率,GDP增长率,总资产,H指数。

图19.45　模型汇总

如图19.46所示,模型总的检验结果提示,整个模型用于预测因变量是有价值的,具有统计学意义。

模型		平方和	df	均方	F	Sig.
1	回归	1.376	7	.197	6.844	.000[b]
	残差	2.155	75	.029		
	总计	3.531	82			

a. 因变量:净资产收益率。

b. 预测变量:(常量),证券市场增长率,ACR,市场份额,经营效率,GDP增长率,总资产,H指数。

图19.46　Anova[a]

续表

函数形式	返回值类型	函数说明
DATE. WKYR(weeknum, year)	日期数值型	返回日期:"year"年第"weeknum"周第1天距基线日期的秒数 例:DATE. WKYR(21,1982)意思是1982年第21周
DATE. YRDAY(year,daynum)	日期数值型	返回日期:"year"年第"daynum"天距基线日期的秒数 例:DATE. YRDAY(1982,21)意思是1982年第21天
XDATE. DATE(datevalue)	数值型	返回"datevalue"距离1582年1月1日的秒数,与CTIME. SECONDS(timevalue)相似 例:XDATE. DATE(1952/02/03) =11654150400.00
XDATE. HOUR(datevalue)	数值型	返回"datevalue"为本天第几时,为整数,介于0~23之间 例:XDATE. HOUR(02 - MAR - 2003 02:30:30) =2
XDATE. JDAY(datevalue)	数值型	返回"datevalue"为本年度第几天,为整数,介于1~366之间 例:XDATE. JDAY(31 - DEC - 2004) =366
XDATE. MDAY(datevalue)	数值型	返回"datevalue"为本月第几天,为整数,介于0~31之间 例:XDATE. MDAY(31 - DEC - 2003) =31
XDATE. MINUTE(datevalue)	数值型	返回"datevalue"为本时第几分,为整数,介于0~59之间 例:XDATE. MINUTE(02 - MAR - 2003 02:30:29) =30
XDATE. MONTH(datevalue)	数值型	返回"datevalue"为本年度第几月,为整数,介于1~12之间 例:XDATE. MINUTE(02 - MAR - 2003 02:30:29) =30
XDATE. QUARTER(datevalue)	数值型	返回"datevalue"为本年度第几季,为整数,介于1~4之间 例:XDATE. QUARTER (02 - MAR - 2003 02:30:29) =1
XDATE. SECOND(datevalue)	数值型	返回"datevalue"为本分第几秒,为整数,介于1~60之间 例:XDATE. SECO - MAR - 2003 02:30:29) =29
XDATE. TDAY(datevalue)	数值型	返回"datevalue" SECONDS(timeval 例:XDATE. TD
XDATE. TIME(datevalue)	日期数值型	返回"datevalu 例:XDATE. T
XDATE. WEEK(datevalue)	数值型	返回"dateva 例:XDATE
XDATE. WKDAY(datevalue)	数值型	返回"dat 例:XDA

与CTIME.

VALUE(variable)　SYSMIS(num　VALID(va　MIS

数,需要两个或两
的最小个数

或两个以上value,可以为