第 3 章　SPSS Modeler 的数据理解*

依据数据挖掘方法论，数据理解在数据挖掘过程中起着举足轻重的作用，其目的是把握数据的总体质量，了解变量取值的大致范围。

数据理解主要包括数据质量评估和调整、数据的有序浏览和多维度汇总等。

SPSS Modeler 的数据准备节点主要分布在节点工具箱的“字段选项”“输出”“记录选项”选项卡中。

3.1　变 量 说 明

变量说明是确保高质量数据的有效途径。变量说明包括两个主要方面：第一，对数据流中变量取值的有效性进行限定、检查和调整；第二，对各个变量在未来数据建模中的角色进行说明。

可通过“字段选项”选项卡中的“类型”节点进行变量说明。

以学生参加某次社会公益活动的数据（文件名为 Students.xlsx）为例，讲解变量说明的具体操作方法。

首先，选择“源”选项卡中的“Excel”节点，添加到数据流编辑区域。建立两个 Excel 节点，分别读入 Students.xlsx 文件中的老生数据（Students）和新生数据（NewStudents）；其次，选择“字段选项”选项卡中的“合并”节点，将其添加到数据流中并分别与两个 Excel 节点相连；最后，选择“输出”选项卡中的“表格”节点浏览数据。发现数据存在以下问题。

- 家庭人均年收入变量，有部分样本观测取值$null$，表示空缺；同时，还有一个样本观测取值为 999999，这里认为这是一个明显错误的数据，应对此进行说明或调整。
- 是否无偿献血变量值填写不规范。规范值应为 Yes 和 No，但有些填写了 1（表示 Yes）和 0（表示 No）。应将 1 替换为 Yes，0 替换为 No。

可利用“类型”节点解决上述问题。选择“字段选项”选项卡中的“类型”节点，将其添加到“合并”节点的后面。右击，选择弹出菜单中的“编辑”进行节点的参数设置。“类型”节点参数设置中的“格式”选项卡内容比较简单，这里只重点讨论其中的“类型”选项卡，如图 3.1 所示。

“类型”选项卡以列表形式依次给出了各变量的变量名（字段）、计量类型（测量）、取值范围（值）、缺失值（缺失）、检查及变量角色（角色）。这与第 2 章数据读入相关节点的“类型”选项卡完全相同。

电子工业出版社有限公司
版权所有

*本章的数据流文件：数据理解.str。

图 3.1 “类型”节点的“类型”选项卡

需要说明的是，SPSS Modeler 单独设置“类型”节点的意义在于数据流中的数据是静态的。通常，数据源节点中的数据可能会更新，数据流也会派生出一些新的变量，或数据流进行了数据的集成操作，或对原有变量的类型进行了调整。如果没有执行整个数据流，也就是说，新数据没有“流过”每个节点，那么变量属性还会保持原有的“静态”，即变量的取值范围、类型等是不会动态调整的。“类型”节点允许用户及时跟踪数据的动态变化，使整个数据流的数据保持一致。“类型”节点可设置在数据流的任何恰当位置上。

3.1.1 变量的重新实例化

数据读入时变量需要进行实例化。当数据源节点中的数据有更新，数据流派生出一些新的变量，进行了数据集成操作或原有变量的类型有了调整时，变量需要重新实例化。

“类型”选项卡中的“读取值”“清除值”“清除所有值”3 个按钮可用于变量的重新实例化。具体操作步骤如下。

- 第一步，单击“清除值”或“清除所有值”按钮，强制将所有变量变为非实例化状态。于是，所有变量的“值”列自动取值为“<读取>”。从效率角度考虑，如果并非所有变量都需重新实例化，则在需要重新实例化的变量行的“值”下拉框中，选择“<读取>”或“<读取+>”即可，如图 3.2 所示。

“<读取>”表示读入数据重新实例化；“<读取+>”表示读入数据且新数据自动追加到原有数据的后面；“<传递>”表示不读入变量值；“<当前>”表示保持变量的当前值，不重新实例化，是默认选项。

- 第二步，单击“读取值”按钮进行变量的重新实例化。于是，“值”列将显示各变量值的取值范围。

电子工业出版社有限公司 版权所有 盗版必究

图 3.2 “类型”节点的变量重新实例化

3.1.2 有效变量值和无效值调整

有效变量值是变量正常取值范围内的值，无效值是变量有效取值范围之外的值，通常称为缺失值。SPSS Modeler 中的缺失值通常包括两类：一类是系统缺失值，用$null$表示，还包括空串和空格等；另一类是用户缺失值，主要指那些取值明显不合理的数据。

变量有效取值范围和缺失值的说明应通过选择“缺失”列的选项来实现。其中：

- 开(*)：表示允许相应变量取系统缺失值和用户缺失值，且不进行调整。
- 关：表示不允许相应变量取缺失值。
- 指定：说明变量的有效取值范围等，并指定数据调整方法。

对是否无偿献血变量的说明步骤为：首先，在相应行的“缺失”列中，选择“指定”，窗口如图 3.3 所示；其次，删除值列中的 0 和 1。添加 No 和 Yes 的值标签，说明是否无偿献血的规范取值；最后，指定是否无偿献血变量的用户缺失值 0、1。

图 3.3　变量说明的“指定”窗口（一）

不同类型变量的“指定”窗口略有差别，但含义大体一致。

- 描述：用于输入变量名标签，是变量含义的简短说明文字。

电子工业出版社有限公司 盗版必究 版权所有

- 测量、存储：显示当前变量的计量类型和存储类型。
- 值：指定变量取值范围的确定依据。

其中，“读取数据”表示取决于所读取的外部数据；“传递”表示忽略所读取的外部数据；“指定值和标签”表示人为指定变量取值和值标签。用户可根据当前变量的实际意义，指定其合理的取值，并在“标签”列中输入关于变量值含义的简短说明文字。

本例中，为说明是否无偿献血的规范取值，分别在 No 和 Yes 对应的“标签”列中输入值标签“未曾无偿献血”“曾无偿献血”。

- 检查值：选择指定对变量不合理值的调整方法。

其中，“无”表示不进行调整；“无效”表示将用户缺失值调整为系统缺失值$null$；“强制”表示调整为 SPSS Modeler 默认的指定值；“丢弃”表示剔除相应样本数据；“警告”表示遇到不合理取值时给出警告信息；“终止”表示遇到不合理取值时终止数据流的执行。

- 定义空白：选中该选项，表示视“缺失值”表中的值及某区间内的连续值、$null$、空格为空。

其中，在“缺失值”表中输入若干个离散值；在“范围”输入连续区间；“空”和“空白”分别表示$null$和空格。所指定的值均当空处理。

指定为空的目的是将无须或无法调整的用户缺失值和系统缺失值，与变量的正常取值区别开，便于后续的数据分析。选中该项后，图 3.1 窗口相应变量的“缺失”列中将自动显示星号（*），表示相应变量存在用户缺失值和系统缺失值，即便指定数据调整方法，也不进行调整。

本例中，指定字符 0 和 1 为空。不调整 0 和 1 的原因在于 SPSS Modeler 提供的调整策略不适合该问题。该问题将在第 4 章进行讨论。

需要说明的是，SPSS Modeler 此处的空，并非一般意义上的空串，它可以是数值，也可以是$null$，还可以是空格。

对变量家庭人均年收入的说明步骤为：首先，在相应行的“缺失”列中选择“指定”，窗口如图 3.4 所示；其次，指定变量值的调整方法。

图 3.4　变量说明的“指定”窗口（二）

工业出版社有限公司 盗版必究 版权所有

本例中，家庭人均年收入的取值范围不能直接由外部数据决定，否则 SPSS Modeler 将视 999999（用户缺失值）为正常值。应在“下限”和“上限”框中手工输入合理的取值区间为 6617.0～503308.0。同时，由于希望对家庭人均年收入中的 999999 和$null$值进行调整，因此，不应选中“定义空白”选项，SPSS Modeler 将自动视 999999 和$null$为超出取值范围的不合理取值，并按“强制”方法进行调整。返回图 3.1 窗口后，家庭人均年收入的“缺失”列中为空，表示该变量不存在缺失值。

3.1.3　变量角色的说明

变量角色是指变量在模型建立时的角色。变量的角色不同，其作用也不同。

模型建立时，有些变量用于解释其他变量，称为解释变量或自变量，SPSS Modeler 称之为输入变量。有的变量需被其他变量解释，称为被解释变量或因变量，SPSS Modeler 称之为目标变量。

例如，在分析顾客的收入对其消费的影响时，收入就是输入变量，消费就是目标变量。

变量角色的说明可通过图 3.1 中的“角色”列指定，如图 3.5 所示。

图 3.5　变量角色说明

除此之外，SPSS Modeler 还将变量角色进行了拓展，具体如下。

- 任意：在某些模型中，有的变量既可作为输入变量也可作为目标角色。

例如，在根据顾客的收入和消费数据将客户划分为不同客户群的分析中，收入和消费既是输入变量也是目标变量，为任意角色。

- 分区：样本集分割角色。

样本集分割角色的变量应是只能有两个或三个类别值的分类变量。其中，第一个分类值是训练样本集标记，第二个是测试样本集标记，第三个是验证样本集标记。

- 无：如果某变量不参与数据建模，则可指定它为“无”角色。“无类型”变量自动默认为该角色。

这里，为分析学生是否参加某次社会公益活动受哪些因素的影响，指定学生编号角色为“无”，是否参与为目标变量，其他变量为输入变量。

电子工业出版社有限公司 盗版必究 版权所有

需要说明的是，变量角色说明在数据挖掘的后期建模中才会涉及，通常不在数据理解阶段考虑。相关内容安排于此是为了便于对 SPSS Modeler 软件操作的讲解。

3.2　数据质量的评估和调整

高质量数据是数据分析的前提和可靠分析结论的保障。数据质量的评估和调整是对现有数据的取值异常程度及缺失情况等进行综合评价，并借助统计方法对其进行适当调整和插补。

3.2.1　数据的基本特征与质量评价报告

SPSS Modeler 的数据质量评估主要对数据的缺失、离群点和极端值等情况进行评估。具体包括完整变量比例的计算、完整观测比例的计算，以及其他评价指标的计算等。

可通过“输出”选项卡中的“数据审核”节点评估数据质量。

这里，以一份电信客户的模拟数据为例，该数据为 SPSS 格式，文件名为 Telephone.sav。该数据包括居住地、年龄、婚姻状况、收入、教育水平、性别、家庭人数、开通月数、无线服务、基本费用、上月限制性免费服务项目的费用、无线服务费用、是否电子支付、客户所申请的服务套餐类型、是否流失 15 个变量。利用这份数据，可分析流失客户的一般特征，同时建立模型进行客户流失的预测。本节只对该数据的质量进行考察。

具体操作步骤如下。

- 第一步，建立“SPSS 文件”节点读入 Telephone.sav 数据。
- 第二步，建立“类型”节点说明变量角色。这里，指定是否流失为目标变量，其他变量均为输入变量。指定收入小于 20 的值为用户缺失值。
- 第三步，选择“输出”选项卡中的“数据审核”节点，将其添加到数据流的相应位置上。右击，选择弹出菜单中的“编辑”进行节点的参数设置。

“数据审核”节点的参数设置包括“设置”“质量”“输出”等选项卡。

1.“设置”选项卡

“设置”选项卡用于指定质量探索的变量，以及计算输出哪些统计指标，如图 3.6 所示。其中：

- 缺省：表示评估节点包含的所有变量的质量，并且默认数据流前面“类型”节点指定的目标变量将为“交叠字段”。如果交叠字段（变量）为分类型变量，则所绘统计图用于反映交叠变量不同取值下其他变量的分布特征。如果交叠变量为数值型变量，则将计算交叠变量与其他变量的简单相关系数[①]、相关系数 *t* 检验的观测值和自由度、概率 *P*–值及协方差等。

① 简单相关系数：反映两数值型变量线性相关程度的统计指标，取值为–1～+1。大于 0 表示两变量存在正的线性相关关系，小于 0 表示两变量存在负的线性相关关系。绝对值大于 0.8 表示两变量之间具有较强的线性关系，绝对值小于 0.3 表示两变量之间的线性相关关系较弱。

电子工业出版社有限公司 版权所有 盗版必究

- 使用定制字段：表示用户自行指定对哪些变量的质量进行评估。同时，如有必要，还可在“交叠字段”框中指定交叠变量。
- 显示：“图形”主要包括柱形图、直方图和散点图（当交叠变量为数值型时）；“基本统计量”主要包括数值型变量的最小值、最大值、均值、标准差、偏态系数[①]等；“高级统计量”主要包括总和、极差、均值标准误差、方差、峰度系数[②]等。

图 3.6 “数据审核”节点的“设置”选项卡

2.“质量”选项卡

“质量”选项卡用于设置反映数据质量的评价指标，以及数据离群点（离群值）和极端值（极值）的诊断标准等，如图 3.7 所示。

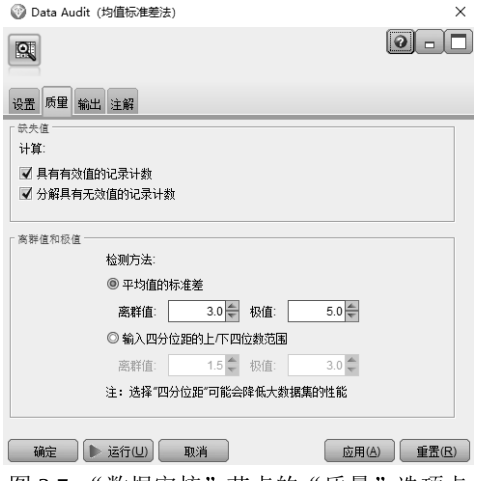

图 3.7 “数据审核”节点的“质量”选项卡

① 偏态系数：反映变量分布对称性的统计指标。偏态系数为 0，表示分布对称；偏态系数大于 0，表示呈右偏不对称分布；偏态系数小于 0，表示呈左偏不对称分布。偏态系数绝对值越大，表示不对称程度越大。

② 峰度系数：反映变量分布陡缓性的统计指标。峰度系数为 0，表示分布陡缓程度同标准正态分布；峰度系数大于 0，表示尖峰分布；峰度系数小于 0，表示平峰分布。

电子工业出版社有限公司 盗版必究 版权所有

其中：

- “缺失值”框：“具有有效值的记录计数”，表示计算各变量的有效样本量；“分解具有无效值的记录计数”，表示计算各变量取各种无效值的样本量。
- “离群值和极值”框：用来指定离群值和极端值的诊断标准。“平均值的标准差”，表示以均值为中心，取值在默认的 3 个标准差以外的变量值为离群值，在默认的 5 个标准差以外的变量值为极值；“输入四分位距的上/下四位数范围”，表示与上四分位数（或下四分位数）的绝对差大于默认的 1.5 倍四分位差[①]时为离群值，大于默认的 3 倍四分位差时为极值。

本例选择“平均值的标准差”方法，且按默认的标准进行诊断。

执行“数据审核”节点，生成的分析表名显示在流管理窗口的“输出”选项卡中。分析表包括“审计”和“质量”两个选项卡。案例的数据审核结果如图 3.8 所示。

图 3.8　案例的数据审核结果

SPSS Modeler 以列表形式依次显示了指定变量的变量名（字段）、图形、测量（计量类型）、相关的描述统计量，以及分类型变量的类别个数（唯一）和有效样本量。

单击窗口工具栏中的按钮，允许用户选择计算其他描述统计量，如峰度系数等；按钮和按钮可用于指定统计图形的显示方向为上下纵向显示（柱形图）还是左右横向显示（条形图）。单击各列的标题处，可按相应列重排数据审核结果。图中深色部分表示目标变量（是否流失）取 Yes（流失）的情况。可以看到，流失客户在各变量不同取值上都有分布。例如，图形粗略显示，在开通月数变量上，开通月数比较短的客户其流失比例相对较大，而在其他变量上的分布差异并不十分明显。另外，收入变量呈明显的右偏不对称分布，偏态系数（偏度）高达 6.532。

在数据质量评估中，应重点关注“有效”列。可以看到，在 1000 个样本观测中，收入变

① 四分位差：等于上四分位数－下四分位数。

电子工业出版社有限公司 盗版必究 版权所有

量上仅有 921 个有效值（原因是之前定义了小于 20 的值为用户缺失值）。进一步，观察各个变量的离群点和极值情况。

观察本例输出结果的“质量”选项卡内容，如图 3.9 所示。

字段	测量	离群值	极值	操作	缺失插补	方法	完成百分比	有效记录	空值	字符型空值
居住地	名义	--	--	--	从不	固定	100	1000	0	
年龄	连续	0	0	无	从不	固定	100	1000	0	
婚姻状况	标记	--	--	--	从不	固定	100	1000	0	
收入	连续	9	6	无	从不	固定	92.1	921	0	
教育水平	有序	--	--	--	从不	固定	100	1000	0	
性别	标记	--	--	--	从不	固定	100	1000	0	
家庭人数	连续	6	0	无	从不	固定	100	1000	0	
开通月数	连续	0	0	无	从不	固定	100	1000	0	
无线服务	标记	--	--	--	从不	固定	100	1000	0	
基本费用	连续	18	4	无	从不	固定	100	1000	0	
免费部分	连续	9	1	无	从不	固定	100	1000	0	
无线费用	连续	8	1	无	从不	固定	100	1000	0	
电子支付	标记	--	--	--	从不	固定	100	1000	0	
套餐类型	名义	--	--	--	从不	固定	100	1000	0	
流失	标记	--	--	--	从不	固定	100	1000	0	

图 3.9 “质量”选项卡

SPSS Modeler 以列表形式显示了关于数据质量的评价指标，包括完成百分比（即在该变量上取有效值的样本个数占总样本量的比例）、有效记录（取有效值的样本量）及取各类无效值（如离群值、极值、空值、空白等）的样本量。另外，还列出了对离群值、极值及缺失值的调整方法。结果显示：

- 首先，“完整字段（%）”，即均取有效值的变量占总变量个数的百分比，这里为 93.33%；“完整记录（%）”，即均取有效值的样本量占总样本量的百分比，这里为 92.1%。
- 其次，观察每个变量的情况。例如，收入变量，其上有 9 个离群值和 6 个极值，且其取有效值的为 92.1%；基本费用变量，其上有 18 个离群值和 4 个极值。

3.2.2　变量值的调整

SPSS Modeler 的变量值调整是在“数据审核”节点执行结果的基础上，针对数据中的离群值、极值及缺失值，根据用户选择的方法进行的调整或插补。

1．离群值和极值的调整

SPSS Modeler 对离群值和极值的修正方法列在图 3.9 的“操作”列中。

具体操作步骤如下。

第一步，选中某个变量行。

第二步，下拉相应行的“操作”框选择调整方法，如图 3.10 所示。

“操作”列中提供了以下离群值和极值的调整方法。

- 强制：表示将离群值或极值调整为距它们最近的正常值。

例如，如果离群值定义为 3 个标准差以外的变量值，则所有离群值可用 3 个标准差上的变量值最大值或最小值替代。

电子工业出版社有限公司 盗版必究 版权所有

图 3.10　选择变量值的调整方法

- 丢弃：表示剔除离群值或极值。
- 无效：表示离群值或极值用系统缺失值$null$替代。
- 强制替换离群值/丢弃极值：表示按强制方法修正离群值，并剔除极值。
- 强制替换离群值/使极值无效：表示按强制方法修正离群值，并将极值替换为系统缺失值$null$。

第三步，选中需要调整的变量行，选择窗口菜单中“生成”下的“离群值和极值超节点”，弹出如图 3.11 所示对话框。

图 3.11　调整离群值和极值

“离群值和极值超节点”表示 SPSS Modeler 将自动生成一个包含若干个“选择”和“填充”节点的超节点，用于根据用户指定的调整方法，调整离群值和极值。“选择”节点用于筛

电子工业出版社有限公司 盗版必究 版权所有

选样本观测，“填充”节点用于变量值的重新计算，具体内容将在第 4 章讲解。

在弹出对话框中，选择“仅所选字段”，表示仅调整所选变量中的离群值和极值；选择“所有字段”，表示对所有变量进行调整。

于是，SPSS Modeler 会生成一个超节点并自动放置在数据流编辑区域。用户只需将所生成的超节点连接到数据流的恰当位置上，并再建立一个“数据审核”节点，即可查看变量值调整的效果。但均值标准差的离群点和极端值判断比较适合对称分布。如果不能确定，较适合采用四分位差方法，本例剔除了 66 个样本，且数据质量有明显提高。数据质量评价和变量值调整数据流如图 3.12 所示。

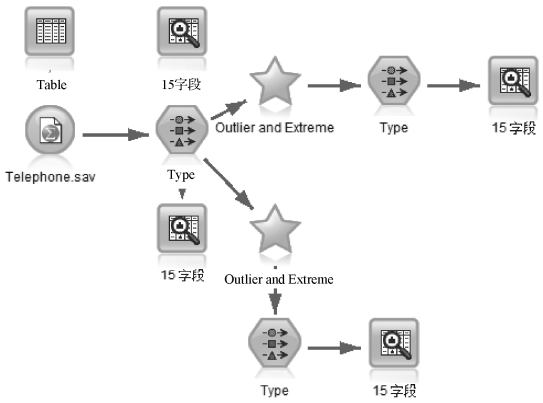

图 3.12　数据质量评价和变量值调整数据流

其中第二个“类型”节点用于重新实例化。

2．缺失值的调整

SPSS Modeler 对缺失值的修正方法列在了图 3.9 的“缺失插补”和“方法”列中。

具体操作步骤如下。

第一步，选中某个变量行。

第二步，下拉相应行的“缺失插补”框选择调整对象。

“缺失插补”列默认值为“从不”，表示不对缺失值做插补处理。下拉“缺失插补”框，可重新指定对哪种情况进行插补，其中：

- 空白值：表示将对空做插补。
- 空值：表示将对系统缺失值$null$做插补。
- 空白值和空值：表示将对空和系统缺失值做插补。
- 条件：表示将对满足指定条件的样本观测中的缺失值做插补，如图 3.13 所示。

图 3.13　对满足指定条件的变量值做插补

首先，在“插补时间[①]”中选择“条件”，并在“条件”框中输入一个 CLEM 条件表达式。CLEM 条件表达式将在后续章节讲解。

① 英文直译，更准确的翻译为：插补时机。

电子工业出版社 版权所有

然后，在“插补方法”中选择采用的插补方法，包括：

- 固定：为默认值，表示插补值为某个固定值。如果选择“固定”方法，还应在“已固定为”的下拉框中选择固定值，可以是平均值、中程数值（即 1/2 的极差）或一个指定的常量。
- 随机：即随机插补，表示插补值为一个服从正态分布或均匀分布的随机值。SPSS Modeler 将给出相应变量的正态分布参数和均匀分布参数，如图 3.14 所示（收入的均值和标准差）。
- 表达式：表示插补值为一个 CLEM 算术表达式的结果。CLEM 算术表达式将在后续章节讲解。
- 算法：表示插补值为模型的预测结果。这里 SPSS Modeler 只给出了 C$RT 分类回归树模型，如图 3.15 所示。分类回归树的原理将在后续章节讲解。

图 3.14　调整为随机数

图 3.15　调整为模型计算结果

于是，图 3.9 中的“缺失插补”和“方法”列将分别显示用户的选择。

第三步，选中需要调整的变量行，选择窗口菜单“生成”下的“缺失值超节点”，弹出如图 3.16 所示对话框。

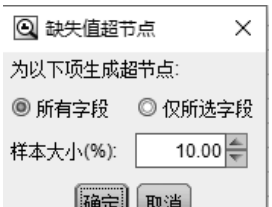

图 3.16　缺失值插补

“缺失值超节点”表示 SPSS Modeler 将自动生成一个包含若干个必要节点的超节点，用于根据用户指定的方法进行缺失值插补。

在弹出对话框中，选择“仅所选字段”，表示仅对所选变量中的缺失值做插补；选择“所有字段”，表示对所有变量的缺失值做插补。在“样本大小(%)”框中给出一个百分比值，默认为 10%，表示对当前样本进行 10%的随机抽样，插补值将基于这 10%的随机样本计算。

于是，SPSS Modeler 会生成一个超节点并自动放置在数据流编辑区域。用户只需将所生成的超节点连接到数据流的恰当位置上并执行即可。

3.2.3　数据质量管理

数据质量管理是指当数据质量评估后，可以将质量不高的变量或样本观测剔除掉，仅保留高质量的变量和样本观测。

1．保留高质量的变量

SPSS Modeler 借助“数据审核”节点的执行结果，可自动保留质量高的变量，剔除质量

电子工业出版社有限公司 版权所有 盗版必究

不高的变量。高质量变量的标准主要指，在该变量上取有效值的样本观测数占总样本量的比例（完整字段%）高于某个指定值。

在图 3.9 所示的窗口中，选择窗口菜单“生成”的“缺失值过滤节点”，弹出如图 3.17 所示的窗口。

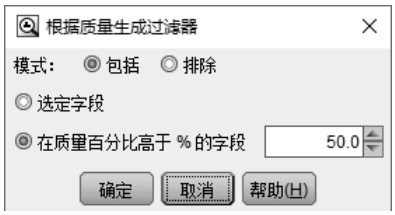

图 3.17 “数据审核”中的“缺失值过滤节点”窗口

其中：

- 模式：“包括”表示保留相应变量；“排除”表示剔除相应变量。
- 选定字段：表示保留或剔除已选定的变量。
- 在质量百分比高于%的字段：给定一个百分比（默认值为 50），表示保留或剔除质量在指定百分比以上的变量。

于是，SPSS Modeler 将在数据流编辑区自动生成一个“过滤器”节点，将节点恰当连接到数据流后，可以看到变量保留或剔除的情况。

例如，如果仅挑出质量在 95%以上的变量，则收入将被自动剔除。SPSS Modeler 生成的“过滤器”节点如图 3.18 所示。

图 3.18 “过滤器”节点

“过滤器”节点用于变量筛选，直观展示变量取舍情况，与第 2 章数据读入节点的“过滤器”选项卡完全相同。用户可以通过鼠标在“过滤器”列上打叉或去掉叉，表示剔除或保留相应变量。另外，SPSS Modeler 的“字段选项”选项卡还提供了专门的“过滤器”节点，其功能与此一致。

2．找出无效样本

SPSS Modeler 借助“数据审核”节点的执行结果，可指定自动找出有效样本观测或无效样本观测。有效样本观测是指那些在指定变量上未取无效值的样本观测，无效样本观测是指那些在指定变量上取了无效值的样本观测。

电子工业出版社有限公司 盗版必究 版权所有

在图 3.9 所示的窗口中，选择窗口菜单“生成”下的“缺失值选择节点”，弹出如图 3.19 所示的窗口。

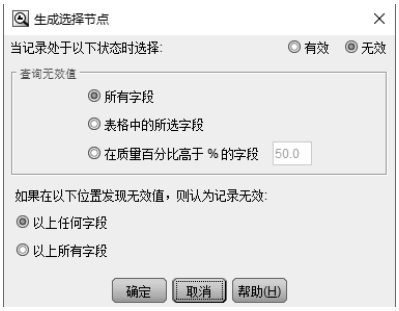

图 3.19 “数据审核”中的“缺失值选择节点”窗口

可以指定找到这样的无效样本观测，即在所有变量、选中变量或质量高于指定百分比的变量上取了无效值，且在前述中的一个变量或所有变量上，取了无效值的样本观测。

需要说明的是，数据质量的评估、数据的调整及根据质量对变量和样本的取舍，是数据挖掘“数据理解”阶段的重要内容。没有高质量的数据基础，就不能有可信的数据分析结论。

3.3　数据的排序

数据排序功能虽然简单，却有广泛的应用，是人们把握数据取值状态的最简捷的途径。

将样本数据按某个或某几个变量值的升序或降序重新排列，一方面便于浏览数据，了解变量取值的大致范围；另一方面，还有助于发现数据可能存在的问题，如离群点或极端值等，因为这些值往往表现最大值或最小值。

“记录选项”选项卡中的“排序”节点可实现数据的排序。

这里，仍以上述电信客户数据（文件名为 Telephone.sav）为例。操作目标有两个：第一，按基本费用的降序排列数据；第二，根据客户最终是否流失，将数据按基本费用的降序排序。

第一个操作目标本质是一个单变量排序问题，第二个操作目标可通过多重排序实现。

3.3.1　单变量排序

单变量排序是只根据一个变量的升序或降序重新排列数据，该变量称为排序变量。

本例的第一个操作目标属单变量排序问题。其中，排序变量为基本费用。

首先，选择“SPSS 文件”节点读入 Telephone.sav 数据；然后，选择“记录选项”选项卡中的“排序”节点，添加到数据流的相应位置上。右击，选择弹出菜单中的“编辑”进行节点的参数设置。参数设置的主要选项卡为“设置”选项卡，用于设置排序变量和排序方式，如图 3.20 所示。

电子工业出版社有限公司 盗版必究 版权所有

SPSS Modeler 以列表形式显示排序变量名（字段）和相应的排序方式（顺序）。

首先，单击按钮，在选择变量对话框中选择排序变量，如图 3.21 所示；然后，在“顺序”列指定数据按升序或降序排序。为查看排序结果，应选择“输出”选项卡中的“表格”节点连接到“排序”节点后面。

图 3.20 “排序”节点的“设置”选项卡

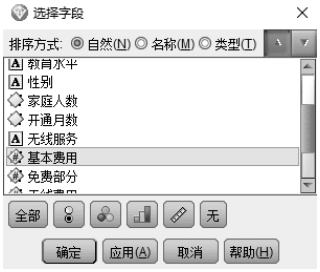

图 3.21　选择变量对话框

图 3.21 所示的选择变量对话框按默认顺序，依次列出了数据流中的所有变量名。可通过选择“排序方式”后的选项改变变量名显示的顺序。其中，“自然”表示按数据流中变量的原有顺序排列；“名称”表示按变量名字母顺序排列；“类型”表示按变量的存储类型排列。单击或按钮，指定变量名排序的升序或降序。选择变量对话框将在 SPSS Modeler 使用中频繁出现，以后不再赘述。

3.3.2　多重排序

多重排序也称多变量排序。应依次指定多个排序变量，分别称为第一排序变量、第二排序变量、第三排序变量，等等。数据排序时，将首先按第一排序变量的升序或降序排序。对第一排序变量取值相同的样本观测，再按第二排序变量的升序或降序排序，依次类推。

本例的第二个操作目标属多重排序问题。其中，第一排序变量为流失，按升序排序；第二排序变量为基本费用，按降序排序，如图 3.22 所示。

于是，数据将首先根据客户是否流失排序，未流失（变量取值为 No）的排在前面，流失（变量取值为 Yes）的排在后面。同时，各类客户内部按基本费用的降序排序。为查看多重排序结果，应选择“输出”选项卡中的“表格”节点连接到“排序”节点后面。

图 3.22　多重排序

需要说明的是，多重排序变量中，“排序方式”框中排在第一行的变量为第一排序变量，排在第二行的为第二排序变量，依次类推。

电子工业出版社有限公司 盗版必究 版权所有

3.4　数据的分类汇总

数据的分类汇总是指，首先，根据所指定的分组变量将数据分成若干组；然后，在各个组内计算汇总变量的基本描述统计量。“记录选项”选项卡中的“汇总”节点可实现数据的分类汇总。

这里，仍以电信客户数据（文件名为 Telephone.sav）为例进行说明。操作目标有两个：第一，分别计算未流失客户和流失客户的基本费用的平均值和标准差；第二，分别针对未流失客户和流失客户群，计算选用不同类套餐类型的客户，其基本费用的平均值和标准差。

第一个操作目标是一个单变量分类汇总问题，第二个操作目标可通过多重分类汇总实现。

3.4.1　单变量分类汇总

单变量分类汇总是只根据一个变量（称为分组变量）对数据分组后，计算其他指定变量（称为汇总变量）的基本描述统计量。

本例的第一个操作目标属单变量分类汇总问题。其中，分组变量为是否流失，汇总变量为基本费用。

首先，选择“SPSS 文件”节点读入 Telephone.sav 数据；然后，选择“记录选项”选项卡中的“汇总”节点，将其添加到数据流的相应位置上。右击，选择弹出菜单中的“编辑”进行节点的参数设置。“设置”是参数设置中的主要选项卡，如图 3.23 所示。

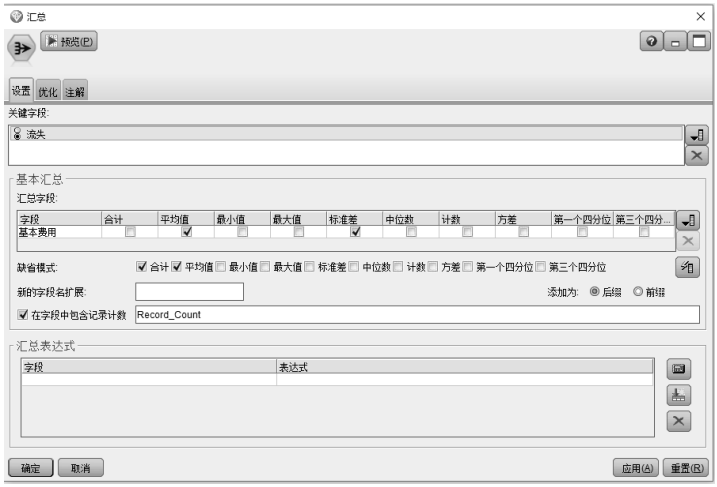

图 3.23　“汇总”节点的“设置”选项卡

其中：

- 在“关键字段”列表中指定分组变量：利用选择变量对话框，选择一个或多个分组变量，该变量通常是分类型变量。
- 在“汇总字段”列表中指定汇总变量：利用选择变量对话框，选择一个或多个汇总变量，并计算基本描述统计量（如均值、标准差等）。
- 在“新的字段名扩展”框中指定扩展名，将作为汇总变量名的后缀或前缀，该变量用

盗版必究
版权所有

来存放分类汇总结果。通常无须指定，汇总结果变量名默认为汇总变量名，后跟下画线和 Sum、Mean 等描述统计量名，如基本费用_Mean。

- 选中“在字段中包含记录计数”，将生成一个默认名为 Record_Count 的变量存放各组的样本量。

为查看分类汇总结果，应选择“输出”选项卡中的“表格”节点连接到“汇总”节点后面。

3.4.2　多重分类汇总

多重分类汇总应依次指定多个分组变量，分别称为第一分组变量、第二分组变量、第三分组变量，等等。分类汇总时，首先将数据按多个分组变量的交叉取值分成若干组；然后再针对各个组，分别计算汇总变量的基本描述统计量。

本例的第二个操作目标属于多重分类汇总问题。其中，第一分组变量为流失，第二分组变量为套餐类型，汇总变量为基本费用，如图 3.24 所示。

图 3.24　多重分类汇总

需要说明的是，多重分类汇总中，“关键字段”框中排在第一行的分组变量为第一分组变量，排在第二行的为第二分组变量，依次类推。

电子工业出版社有限公司
盗版必究
版权所有