IBM SPSS Statistics Base 23

注意

在使用本资料及其支持的产品之前，请阅读第 175 页的『声明』中的信息。

产品信息

本版本适用于 IBM SPSS Statistics V23.0.0 及所有后续发行版和修订版，直到在新版本中另有声明为止。

目录

第 1 章 码本

码本报告活动数据集中所有或指定变量和多响应集的字典信息（如变量名称、变量标签、值标签、缺失值）和汇总统计。对于名义和有序变量以及多响应集，汇总统计包括计数和百分比。对于刻度变量，汇总统计包括平均值、标准差和四分位数。

注意：码本忽略拆分文件状态。这包括为缺失值的多重插补创建的拆分文件组（在缺失值附加选项中可用）。

要获取码本

1. 从菜单中选择:

   分析 > 报告 > 码本
2. 单击“变量”选项卡。
3. 选择一个或多个变量和/或多响应集。

根据需要，您可以:

- 控制显示的变量信息。
- 控制显示的统计（或排除所有汇总统计）。
- 控制变量和多响应集显示的顺序。
- 更改源列表中任何变量的测量级别以更改显示的汇总统计。请参阅 第 3 页的『“码本统计”选项卡』主题以获取更多信息。

更改测量级别

您可以暂时更改变量的测量级别。（您不能更改多响应集的测量级别。它们总是被视为名义变量。）

1. 右键单击源列表中的变量。
2. 从弹出菜单中选择测量级别。

这将暂时更改测量级别。在实际情况下，这仅对数值变量有用。字符串变量的测量级别被限制为名义或有序，二者在“码本”过程中的处理方式相同。

“码本输出”选项卡

“输出”选项卡控制每个变量和多响应集包括的变量信息、变量和多响应集的显示顺序以及可选文件信息表的内容。

变量信息

这控制每个变量显示的字典信息。

位置。代表变量在文件顺序中的位置的整数。这对于多响应集不可用。

标签。与变量或多响应集相关联的描述性标签。

类型。基本数据类型。这可以是*数值*、*字符串*或*多响应集*。

© Copyright IBM Corporation 1989, 2014

格式。变量的显示格式，如 *A4*、*F8.2* 或 *DATE11*。这对于多响应集不可用。

测量级别。可能的值是*名义*、*有序*、*刻度*和*未知*。显示的值是字典中存储的测量级别，不受由更改“变量”选项卡上源变量列表中测量级别所指定的任何临时测量级别覆盖的影响。这对于多响应集不可用。

注意：如果未明确设置测量级别（例如从外部源或新创建的变量读取的数据），那么数值变量的测量级别在第一次数据遍历之前可能是“未知”。请参阅主题以获取更多信息。

角色。某些对话框支持基于定义的角色预先选择分析变量的功能。

值标签。与特定数据值相关联的描述性标签。

- 如果在“统计”选项卡上选择了计数或百分比，那么即使您未在此处选择值标签，输出中仍包括定义的值标签。
- 对于多二分集，“值标签”是集中基本变量的变量标签还是已计算值的标签，这取决于集的定义方式。请参阅主题以获取更多信息。

缺失值。用户定义的缺失值。如果在“统计”选项卡上选择了“计数”或“百分比”，那么定义的值标签将包括在输出中，即使您未在此处选择“缺失值”也是如此。这对于多响应集不可用。

定制属性。用户定义的定制变量属性。对于任何与每个变量相关联的定制变量属性，输出都包括名称和值。请参阅主题以获取更多信息。这对于多响应集不可用。

保留属性。保留系统变量属性。您可以显示系统属性，但是您不得改变这些属性。系统属性名称以美元符号 ($) 开头。不包括名称以对于任何与每个变量相关联的系统属性，输出都包括名称和值。这对于多响应集不可用。

文件信息

可选文件信息表可以包括任何以下文件属性:

文件名。IBM® SPSS® Statistics 数据文件的名称。如果数据集从未以 IBM SPSS Statistics 格式保存，那么就没有数据文件名。（如果在“数据编辑器”窗口的标题栏中没有显示文件名，那么活动数据集没有文件名。）

位置。IBM SPSS Statistics 数据文件的目录（文件夹）位置。如果数据集从未以 IBM SPSS Statistics 格式保存，那么就没有位置。

个案数。活动数据集中的个案个数。这是个案的总数，包括任何由于过滤条件而从汇总统计中排除的个案。

标签。这是由 `FILE LABEL` 命令定义的文件标签（如有）。

文档。数据文件文档文本。

权重状态。如果采用加权，那么显示加权变量的名称。请参阅主题以获取更多信息。

定制属性。用户定义的定制数据文件属性。使用 `DATAFILE ATTRIBUTE` 命令定义的数据文件属性。

保留属性。保留系统数据文件属性。您可以显示系统属性，但是您不得改变这些属性。系统属性名称以美元符号 ($) 开头。不包括名称以对于任何系统数据文件属性，输出都包括名称和值。

变量显示顺序

可使用以下选项来控制变量和多响应集的显示顺序。

依字母顺序排列。依变量名称的字母顺序。

文件。变量在数据集中的显示顺序（变量在数据编辑器中的显示顺序）。在升序方式中，多响应集最后显示（在所有选定变量之后）。

测量级别。按测量级别排序。这将创建四个排序组：名义、有序、刻度和未知。多响应集被视为名义。

注意：如果未明确设置测量级别（例如从外部源或新创建的变量读取的数据），那么数值变量的测量级别在第一次数据遍历之前可能是“未知”。

变量列表。变量和多响应集在“变量”选项卡上的选定变量列表中显示的顺序。

定制属性名称。排序顺序选项列表同时还包括任何用户定义的定制变量属性的名称。在升序方式中，没有属性的变量排在顶端，接着是有属性但尚未定义值的变量，然后是已为属性定义值的变量，这些都按值的字母顺序排列。

最大类别数

如果输出包括每个唯一值的值标签、计数或百分比，那么如果值的数量超过指定的值，您可以在表中不显示此信息。缺省情况下，如果变量唯一值的数量超过 200，那么不显示此信息。

“码本统计”选项卡

“统计”选项卡允许您控制输出中包括的汇总统计，或不显示整个汇总统计。

计数和百分比

对于名义和有序变量、多响应集以及刻度变量的标签值，可用的统计是：

计数。有变量的每个值（或值范围）的个案的计数或个数。

百分比 (Percent). 具有特定值的个案的百分比。

集中趋势和离散

对于刻度变量，可用的统计是：

平均值 (Mean). 集中趋势的测量。算术平均，总和除以个案个数。

标准差 (Standard Deviation). 对围绕平均值的离差的测量。在正态分布中，68% 的个案在平均值的一倍标准差范围内，95% 的个案在平均值的两倍标准差范围内。例如，在正态分布中，如果平均年龄为 45，标准差为 10，那么 95% 的个案将处于 25 到 65 之间。

四分位数 (Quartiles). 显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

注意：您可以在“变量”选项卡上的源变量列表中临时更改与变量相关联的测量级别（从而更改针对该变量显示的汇总统计）。

第 2 章 频率

频率过程提供有助于描述多种类型的变量的统计和图形显示。频率过程是查看数据理想的开始位置。

对于频率报告和条形图，可以用升序或降序排列不同的值，也可以按其频率对类别进行排序。当变量具有许多不相同的值时，可提取频率报告。您可以使用频率（缺省值）或百分比标记图表。

示例。按行业类型划分的公司客户的分布是什么？从输出中可以了解到客户的 37.5% 来自政府机构，24.9% 来自公司，28.1% 来自学术机构，9.4% 来自保健行业。对于连续的定量数据（例如，销售收入），您会了解到平均产品销售额为 3,576 美元，标准差为 1,078 美元。

统计和图。频率计数、百分比、累积百分比、平均值、中位数、众数、和、标准差、方差、范围、最小值和最大值、平均值标准误差、偏度和峰度（两者都带有标准误差）、四分位数、用户指定的百分位数、条形图、饼图和直方图。

频率数据注意事项

数据。使用数值代码或字符串以对分类变量进行编码（名义或序数级别测量）。

假设。特别对于已排序或未排序的类别的变量，表格和百分比可以提供对所有分布中的数据都有用的描述。大多数可选汇总统计（如平均值和标准差）是基于正态理论的，它们适用于对称分布的定量变量。稳健统计（如中位数、四分位数和百分位数）适合于可能符合或可能不符合正态假设的定量变量。

获取频率表

1. 从菜单中选择:

   分析 > 描述统计 > 频率...
2. 选择一个或多个分类变量或定量变量。

根据需要，您可以:

- 单击**统计**以获得定量变量的描述统计。
- 单击**图表**以获得条形图、饼图和直方图。
- 单击结果显示顺序的**格式**。

频率统计

百分位值。一个定量变量的值，其将排序过的数据分组，以使某个百分比在上而另外一个百分比在下。四分位数（第 25、50、75 个百分位数）将观察值分为四个大小相等的组。如果您希望相等组的数目不等于 4，请选择 **n 个相等组的分割点**。您也可指定单个百分位数（例如，第 95 个百分点，有 95% 的观察值大于该值）。

集中趋势。描述分布位置的统计，包括平均值、中位数、众数和所有值的总和。

- *平均值 (Mean)*. 集中趋势的测量。算术平均，总和除以个案个数。
- *中位数 (Median)*. 第 50 个百分位，大于该值和小于该值的个案数各占一半。如果个案个数为偶数，那么中位数是个案在以升序或降序排列的情况下最中间的两个个案的平均。中位数是集中趋势的测量，但对于远离中心的值不敏感（这与平均值不同，平均值容易受到少数多个非常大或非常小的值的影响）。

© Copyright IBM Corporation 1989, 2014

- *众数*。最常出现的值。如果出现频率最高的值不止一个，那么每一个都是一个众数。“频率”过程仅报告此类多个众数中最小的那个。
- *总和 (Sum)*. 所有带有非缺失值的个案的值的合计或总计。

离差。测量数据中变异和展开的统计，包括标准差、方差、范围、最小值、最大值和平均值标准误差。

- *标准差*。对围绕平均值的离差的测量。在正态分布中，68% 的个案在平均值的一倍标准差范围内，95% 的个案在平均值的两倍标准差范围内。例如，在正态分布中，如果平均年龄为 45，标准差为 10，那么 95% 的个案将处于 25 到 65 之间。
- *方差 (Variance)*. 对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。度量方差的单位是变量本身的单位的平方。
- *范围 (Range)*. 数值变量最大值和最小值之间的差；最大值减去最小值。
- *最小值 (Minimum)*. 数值变量的最小值。
- *最大值 (Maximum)*. 数值变量的最大值。
- *平均值的标准误差*。取自同一分布的样本与样本之间的平均值之差的测量。它可以用来粗略地将观察到的平均值与假设值进行比较（即，如果差与标准误差的比值小于 -2 或大于 +2，那么可以断定两个值不同）。

分布。偏度和峰度是描述分布形状和对称性的统计。这些统计与其标准误差一起显示。

- *偏度 (Skewness)*. 分布的不对称性测量。正态分布是对称的，偏度值为 0。具有显著的正偏度的分布有很长的右尾。具有显著的负偏度的分布有很长的左尾。作为一个指导，当偏度值超过标准误差的两倍时，那么认为不具有对称性。
- *峰度 (Kurtosis)*. 观察值聚集在中点周围的程度的测量。对于正态分布，峰度统计的值为 0。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，leptokurtic 分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，platykurtic 分布的尾部比正态分布的尾部要薄。

值是组中点。如果您的数据中的值是组中点（例如，所有年龄在 30 多岁的人都被编码为 35），那么选择此选项以估计原始未分组的数据的中位数和百分位数。

频率图

图表类型。饼图显示各部分对整体的贡献。饼图的每个分区对应于由单个分组变量定义的组。条形图将不同值或不同类别的计数作为单独的条显示，使您可以直观地比较各个类别。直方图也有条，但它们沿着相等的区间刻度进行绘制。每个条的高度是定量变量在该区间内的值的计数。直方图显示分布的形状、中心和分布。叠加在直方图上的正态曲线有助于您判断数据是否为正态分布。

图表值。对于条形图，可以按频率计数或百分比标记刻度轴。

频率格式

排序方式。可根据数据中的实际值或根据这些值的计数（出现的频率）以升序或降序排列频率表。但是，如果您请求直方图或百分位数，那么频率假定变量是定量数据并以升序显示其值。

多个变量。如果您生成多个变量的统计表，您可在单个表中显示所有变量（**比较变量**），或显示每个变量的独立统计表（**按变量组织输出**）。

排除具有多个类别的表。此选项防止显示具有超过指定数目的值的表。

第 3 章 描述性

“描述”过程为单个表中的若干变量显示单变量汇总统计，并计算标准化值（*z* 得分）。变量可以按其平均值（升序或降序）大小、按字母顺序或按您选择变量的顺序（缺省值）进行排序。

当 *z* 得分被保存时，它们将被添加到数据编辑器的数据中并可为图表、数据列表和分析所用。如果变量以不同的单位（例如，人均国内生产总值和受教育人口百分比）记录的，*z* 得分转换会将变量置于更易于直观比较的常用标度中。

示例。如果您的数据中每个个案都包含数月中每天采集的每个销售人员的日销售总额（例如，Bob、Kim、Brian 各有一个条目），那么“描述”过程可以计算每个职员的平均日销售额，并从高到低排列结果。

统计。样本大小、平均值、最小值、最大值、标准差、方差、范围、合计、平均值标准误差、峰度和偏度及两者的标准误差。

描述数据注意事项

数据。以图形方式显示数值变量中的记录错误、离群值和分布异常之后使用这些数值变量。“描述”过程对大文件（数千个案）特别有效。

假设。大多数可用统计（包括 *z* 得分）都基于正态理论，并适合于对称分布的定量变量（定距或者定比测量级别）。避免类别未排序或偏斜分布的变量。*z* 得分的分布与原数据具有相同的形状，因此，计算 *z* 得分并不是排除问题数据的方法。

获取描述统计

1. 从菜单中选择:

   分析 > 描述统计 > 描述性...

2. 选择一个或多个变量。

根据需要，您可以:

- 选择**将标准化得分另存为变量**以将 *z* 得分保存为新变量。
- 单击**选项**选择可选统计和显示顺序。

描述: 选项

平均值与总和。缺省情况下显示平均值（或算术平均数）。

离差。测量数据中的分布或变动的统计包括标准差、方差、范围、最小值、最大值和平均值标准误差。

- *标准差*。对围绕平均值的离差的测量。在正态分布中，68% 的个案在平均值的一倍标准差范围内，95% 的个案在平均值的两倍标准差范围内。例如，在正态分布中，如果平均年龄为 45，标准差为 10，那么 95% 的个案将处于 25 到 65 之间。
- *方差 (Variance)*. 对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。度量方差的单位是变量本身的单位的平方。
- *范围 (Range)*. 数值变量最大值和最小值之间的差；最大值减去最小值。
- *最小值 (Minimum)*. 数值变量的最小值。

© Copyright IBM Corporation 1989, 2014

- *最大值 (Maximum).* 数值变量的最大值。
- *平均值的标准误差.* 取自同一分布的样本与样本之间的平均值之差的测量。它可以用来粗略地将观察到的平均值与假设值进行比较（即，如果差与标准误差的比值小于 -2 或大于 +2，那么可以断定两个值不同）。

分布。峰度和偏度是描绘分布形状和对称情况的统计。这些统计与其标准误差一起显示。

- *峰度 (Kurtosis).* 观察值聚集在中点周围的程度的测量。对于正态分布，峰度统计的值为 0。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，leptokurtic 分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，platykurtic 分布的尾部比正态分布的尾部要薄。
- *偏度 (Skewness).* 分布的不对称性测量。正态分布是对称的，偏度值为 0。具有显著的正偏度的分布有很长的右尾。具有显著的负偏度的分布有很长的左尾。作为一个指导，当偏度值超过标准误差的两倍时，那么认为不具有对称性。

显示顺序。缺省情况下，将按您选择变量的顺序显示变量。（可选）您可以按字母顺序升序或降序显示变量。

DESCRIPTIVES 命令的附加功能

使用命令语法语言还可以:

- 保存某些变量而不是所有变量的标准化得分（*z* 得分）（使用 VARIABLES 子命令）。
- 指定包含标准化得分的新变量的名称（使用 VARIABLES 子命令）。
- 从分析中排除任意变量含缺失值的个案（使用 MISSING 子命令）。
- 按照任何统计的值，而不仅是平均值对显示中的变量进行排序（使用 SORT 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 4 章 探索

“探索”过程既可以为所有个案也可以分别为个案组生成汇总统计和图形显示。使用“探索”过程有很多原因：数据过滤、离群值识别、描述、假设检验以及描述子群体（个案组）之间差异的特征。通过数据过滤可以得知您具有异常值、极值、数据中的缺口或其他特性。探索数据可以帮助确定您正考虑用于数据分析的统计方法是否合适。如果该方法要求数据呈正态分布，您可能通过探索得知需要进行转换数据。或者，您可能认为需要进行非参数检验。

示例。看一下老鼠在四种不同强化时制下的迷宫学习时间的分布。对于四个组中的每一个，可以发现时间是否近似呈正态分布，以及四个方差是否相等。您也可以标识具有 5 个最长时间和 5 个最短时间的个案。箱图和茎叶图以图形方式汇总每个组的学习时间的分布。

统计和图。平均值、中位数、5% 切尾平均值、标准误差、方差、标准差、最小值、最大值、范围、四分位距、偏度和峰度及它们的标准误差、平均值的置信区间（和指定的置信度）、百分位、Huber 的 M 估计、Andrews 波估计量、Hampel 的重新下降 M 估计和 Tukey 的双权重估计量、五个最大值和五个最小值、带用于检验正态性的 Lilliefors 显著性水平的 Kolmogorov-Smirnov 统计和 Shapiro-Wilk 统计。箱图、茎叶图、直方图、正态图、带 Levene 检验和转换的分布-水平图。

探索数据注意事项

数据。“探索”过程可用于定量变量（定距或者定比级别度量）。因子变量（用于将数据分为个案组）应具有合理数量的不相同的值（类别）。这些值可以是短字符串或数值。用于在箱图中标记离群值的个案标签变量可以是短字符串、长字符串（前 15 个字节）或数值。

假设。数据不必呈对称或正态分布。

探索数据

1. 从菜单中选择:

   分析 > 描述统计 > 探索...

2. 选择一个或多个因变量。

根据需要，您可以:

- 选择一个或多个因子变量，其值将定义个案组。
- 选择标识变量用于标记个案。
- 单击**统计**以获得稳健估计量、离群值、百分位和频率表。
- 单击**图**以获得直方图、正态概率图和检验以及带 Levene 统计的分布-水平图。
- 单击**选项**以处理缺失值。

探索: 统计

描述性。缺省情况下显示集中趋势度量和离差测量。集中趋势的测量表示分布的位置；包括平均值、中位数、5% 切尾平均值。离差测量显示值的非相似性；包括标准误差、方差、标准差、最小值、最大值、范围、四分位距。描述统计还包括分布形状的测量；偏度和峰度与它们的误差一起显示。还显示平均值的 95% 水平置信区间；您可指定其他置信度。

© Copyright IBM Corporation 1989, 2014

M 估计。样本平均值和中位数的稳健替代值，用于估计位置。计算出的估计量应用到个案的权重不同。显示 Huber 的 M 估计、Andrews 波估计量、Hampel 的重新下降 M 估计和 Tukey 的双权重估计量。

离群值。显示五个最大值和五个最小值（带个案标签）。

百分位数。显示第 5 个、第 10 个、第 25 个、第 50 个、第 75 个、第 90 个和第 95 个百分位的值。

探索: 图

箱图。当您具有一个或多个因变量时，这些选项控制箱图的显示。**按因子级别分组**为每个因变量生成单独的显示。在一个显示中，将为因子变量定义的每个组显示箱图。**不分组**为因子变量定义的每个组生成单独的显示。在一个显示中，为每个因变量并排显示箱图。当不同的变量代表在不同的时间度量的同一个特征时，此显示尤其有用。

描述性。使用“描述”组可以选择茎叶图和直方图。

带检验的正态图。显示正态概率和反趋势正态概率图。显示带用于检验正态性的 Lilliefors 显著性水平的 Kolmogorov-Smirnov 统计。如果指定的是非整数权重，那么在加权样本大小位于 3 和 50 之间时，计算 Shapiro-Wilk 统计。对于无权重或整数权重，在加权样本大小位于 3 和 5,000 之间时，计算该统计。

带 Levene 检验的分布-水平图。控制分布-水平图的数据转换。对于所有分布-水平图，显示回归线的斜率和 Levene 的稳健的方差同质性检验。如果选择转换，那么 Levene 检验基于转换后的数据。如果未选择因子变量，那么不生成分布-水平图。**幂估计**针对所有单元格的中位数的自然对数以及幂转换的估计值生成四分位距的自然对数图，以在各单元格中得到相等的方差。分布-水平图协助确定稳定（使之更相等）组之间方差所需的转换的幂。使用**已变换**可以选择幂替代值之一（可能按幂估计中的推荐），并生成转换数据图。绘制转换数据的四分位距和中位数。**未变换**生成原始数据的图。这等于幂为 1 的转换。

探索: 幂转换

这些是分布-水平图的幂转换。要转换数据，您必须选择转换的幂。您可以选择以下选项之一:

- **自然对数**。自然对数转换。这是缺省值。
- **1/平方根**。对于每个数据值，计算平方根的倒数。
- **倒数**。计算每个数据值的倒数。
- **平方根**。计算每个数据值的平方根。
- **平方**。每个数据值的平方。
- **多维数据集**。每个数据值的多维数据集。

探索: 选项

缺失值。控制对缺失值的处理。

- **按列表排除个案**。从所有分析中排除任何因变量或因子变量具有缺失值的个案。这是缺省值。
- **按对排除个案**。在该组的分析中包含组（单元格）中变量不具有缺失值的个案。该个案可能在其他组中使用的变量中有缺失值。
- **报告值**。因子变量的缺失值被视为单独的类别。为此附加类别生成所有输出。频率表包含缺失值的类别。因子变量的缺失值包含在内，但被标记为缺失。

EXAMINE 命令的附加功能

“探索”过程使用 EXAMINE 命令语法。使用命令语法语言还可以:

- 除由因子变量定义（用 TOTAL 子命令）的组的输出和图之外，还请求合计输出和图。
- 指定一组箱图的常用刻度（用 SCALE 子命令）。
- 指定因子变量的交互作用（用 VARIABLES 子命令）。
- 指定缺省值以外的百分位（用 PERCENTILES 子命令）。
- 根据五种方法中的任意一种计算百分位（用 PERCENTILES 子命令）。
- 指定分布-水平图的任意幂转换（用 PLOT 子命令）。
- 指定要显示的极值的数量（用 STATISTICS 子命令）。
- 指定位置的 M 估计和稳健估计量的参数（用 MESTIMATORS 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 5 章 交叉表

交叉表过程形成二阶和多阶表，并提供了各种双向表检验和相关性测量。表的结构以及类别是否排序决定了要使用的检验或度量。

仅对双向表计算交叉表统计和相关性测量。如果指定一行、一列和一个层因子（控制变量），交叉表过程将为层因子（或两个或更多控制变量的值组合）的每个值形成一个关联统计和量度面板。例如，如果*性别*是一个*已婚*（是、否）与*生活*（生活充满激情、循规蹈矩或索然无味）相对照的表的层因子，那么女性的双向表结果将与男性的双向表结果分开计算，并打印成互相接续的面板格式。

示例。在进行服务（例如：训练和咨询）销售时，小公司的客户与来自较大公司的客户相比，是否可能更具盈利性？您可能从交叉制表中发现，大多数小公司（员工少于 500 人）获得很高的服务利润，而从大多数大公司（员工多于 2,500 人）却获得很低的服务利润。

统计和相关性测量。Pearson 卡方、似然比卡方、线性关联检验、Fisher 精确检验、Yates 校正卡方、Pearson *r*、Spearman Rho、列联系数、Phi、Cramér *V*、对称和非对称 Lambda、Goodman 和 Kruskal tau、不确定性系数、伽玛、Somers *d*、Kendall tau-*b*、Kendall tau-*c*、eta 系数、科恩 Kappa、相对风险估计、几率比、McNemar 检验、Cochran 和 Mantel-Haenszel 统计以及列比例统计。

交叉表数据注意事项

数据。要定义每个表变量的类别，请使用数值或字符串（八个或八个以下字节）变量的值。例如，对于 *gender*，您可用将数据编码为 1 和 2，或编码为 *male* 和 *female*。

假设。如有关统计一节中所述，某些统计和度量假定已排序的类别（有序数据）或数量值（定距或者定比数据）。另有一些统计则在表变量具有未排序的类别（名义数据）时有效。对于基于卡方的统计（Phi、Cramér *V* 和列联系数），数据应该是来自多项分布的随机样本。

注：有序变量可以是代表类别的数值代码（例如：1 = *low*、2 = *medium*、3 = *high*），也可以是字符串值。不过，字符串值的字母顺序将假定反映了类别的真实顺序。例如，对于具有 *low*、*medium*、*high* 值的字符串变量，类别的顺序将解释为 *high*、*low*、*medium*，这个顺序是错误的。通常，使用数值代码代表有序数据更为可靠。

获取交叉制表

1. 从菜单中选择:

   分析 > 描述统计 > 交叉表...

2. 选择一个或多个行变量和一个或多个列变量。

根据需要，您可以:

- 选择一个或多个控制变量。
- 单击**统计**以获取双向表或子表的检验和相关性测量。
- 单击**单元格**以获取观察值和期望值、百分比值和残差。
- 单击**格式**以控制类别的顺序。

© Copyright IBM Corporation 1989, 2014

交叉表: 层

如果选择一个或多个层变量，那么将对每个层变量（控制变量）的每个类别产生单独的交叉制表。例如，如果有一个行变量、一个列变量和一个具有两个类别的层变量，那么可为层变量的每个类别生成一个双向表。要形成另一层控制变量，请单击**下一个**。为每个第一层变量与每个第二层变量（等等）的每种类别组合生成子表。如果请求了统计和相关性测量，那么它们仅应用于双向子表。

交叉表复式条形图

显示复式条形图。复式条形图可帮助汇总个案组的数据。对于在“行”下指定的变量的每个值，均有一个复式条形图。定义每个聚类内的条形图的变量就是您在“列”下指定的变量。对于此变量的每个值，均有一组不同颜色或图案的条形图。如果您在“列”或“行”下指定多个变量，那么为每个双变量组合生成一个复式条形图。

在表层中显示层变量的交叉表

在表层中显示层变量。您可以选择在交叉表中将层变量（控制变量）显示为表层。这允许您创建视图来显示行和列变量的整体统计，以及允许深入层变量的类别。

下面显示了使用数据文件 *demo.sav*（位于安装目录下的 Samples 目录中）的示例及其获取方式:

1. 选择 *Income category in thousands (inccat)* 作为行变量，*Owns PDA (ownpda)* 作为列变量以及 *Level of Education (ed)* 作为层变量。
2. 选择**在表层中显示层变量**。
3. 在“单元格显示”子对话框中选择**列**。
4. 运行“交叉表”过程，双击交叉表并从 Level of Education 下拉列表选择 **College degree**。

交叉表的选定视图显示拥有大学学历的响应者的统计。

交叉表统计

卡方。对于两行两列的表，请选择**卡方**以计算 Pearson 卡方、似然比卡方、Fisher 的精确检验和 Yates 修正卡方（连续性修正）。对于 2×2 表，如果表并非源自于包含期望频率小于 5 的单元格的较大表中的缺失行或缺失列，那么计算 Fisher 精确检验。对于所有其他 2×2 表，计算 Yates 修正卡方。对于具有任意行列数的表，选择**卡方**来计算 Pearson 卡方和似然比卡方。当两个表变量都是定量变量时，**卡方**将产生线性关联检验。

相关性。对于行和列都包含排序值的表，**相关**将生成 Spearman 相关系数 rho（仅数值数据）。Spearman 的 rho 是等级顺序之间的相关性测量。当两个表变量（因子）都是定量变量时，**相关**产生 Pearson 相关性系数 *r*，这是变量之间的线性相关性测量。

名义。对于名义数据（无内在顺序，例如天主教、新教和犹太教），您可以选择**列联系数**、**Phi**（系数）以及 **Cramér V**、**Lambda**（对称和非对称 Lambda 以及 Goodman 和 Kruskal tau）和**不确定性系数**。

- *列联系数*。基于卡方统计的相关性测量。值的范围在 0 到 1 之间，其中 0 表示行变量和列变量之间不相关，而接近 1 的值表示变量之间的相关度很高。可能的极大值取决于表中的行数和列数。
- *Phi 和 Cramer V (Phi and Cramer's V)*. Phi 是基于卡方统计的相关性测量，它将卡方检验统计除以样本大小，并取结果的平方根。Cramer V 是基于卡方统计的相关性测量。
- *Lambda*. 一种相关性测量，它反映使用自变量的值来预测因变量的值时，误差成比例缩小。值为 1 表示自变量能完全预测因变量。值为 0 表示自变量对于预测因变量没有帮助。

- *不确定性系数*。一种相关性测量，它表示当一个变量的值用来预测其他变量的值时，误差成比例下降的程度。例如，值 0.83 指示如果知道一个变量的值，那么在预测其他变量的值时会将误差减少 83%。程序同时计算不确定性系数的对称版本和不对称版本。

有序。对于行和列都包含已排序值的表，请选择**伽玛**（对于 2 阶表，为零阶；对于 3 阶到 10 阶表，为条件）、**Kendall 的 tau-b** 和 **Kendall 的 tau-c**。要根据行类别预测列类别，请选择 **Somers 的 d**。

- *伽玛 (Gamma)*. 两个有序变量之间的对称相关性测量，它的范围是从 -1 到 1。绝对值接近 1 的值表示两个变量之间存在紧密的关系。接近 0 的值表示关系较弱或者没有关系。对于双向表，显示零阶伽玛。对于三阶表到 n 阶表，显示条件伽玛。
- *Somers' d*. 两个有序变量之间相关性测量，它的范围是从 -1 到 1。绝对值接近 1 的值表示两个变量之间存在紧密的关系，值接近 0 则表示两个变量之间关系很弱或没有关系。Somers 的 d 是伽玛的不对称扩展，不同之处仅在于它包含了未约束到自变量上的成对的数目。还将计算此统计的对称版本。
- *Kendall 的 tau-b (Kendall's tau-b)*. 将结考虑在内的有序变量或排序变量的非参数相关性测量。系数的符号指示关系的方向，绝对值指示强度，绝对值越大则表示关系强度越高。可能的取值范围是从 -1 到 1，但 -1 或 +1 值只能从正方表中取得。
- *Kendall's tau-c (Kendall's tau-c)*. 忽略结的有序变量的非参数相关性测量。系数的符号指示关系的方向，绝对值指示强度，绝对值越大则表示关系强度越高。可能的取值范围是从 -1 到 1，但 -1 或 +1 值只能从正方表中取得。

按区间标定。当一个变量为分类变量，而另一个变量为定量变量时，请选择 **Eta**。分类变量必须进行数值编码。

- *Eta*。范围在 0 到 1 之间的相关性测量，其中 0 值表示行变量和列变量之间无相关性，接近 1 的值表示高度相关。Eta 适用于在区间刻度上度量的因变量（例如收入）以及具有有限类别的自变量（例如性别）。计算两个 eta 值：一个将行变量视为区间变量，另一个将列变量视为区间变量。

Kappa. 当两个评分者在估计同一个对象时，科恩 Kappa 度量两者的估计之间的一致性。值为 1 表示完全一致。值为 0 表示几乎完全不一致。Kappa 基于一个正方表，其中的行值和列值表示同一个刻度。任何对一个变量具有观察值但对另一变量不具有观察值的单元格都被赋予计数 0。如果两个变量的数据存储类型（字符串或数字）不相同，那么不计算 Kappa。对于字符串变量，两个变量必须具有相同的定义长度。

风险。对于 2 x 2 表，某因子的存在与某事件的发生之间相关性强度的测量。如果该统计的置信区间包含值 1，那么不能假设因子与事件相关。当因子出现很少时，几率比可用作估计或相对风险。

McNemar。两个相关二分变量的非参数检验。使用卡方分布检验响应改变。"之前与之后"设计中的试验干预会导致响因变量发生变化，它对于检测到这些变化很有用。对于较大的正方表，会报告对称性的 McNemar-Bowker 检验。

Cochran's and Mantel-Haenszel 统计 (Cochran's and Mantel-Haenszel statistics). Cochran 和 Mantel-Haenszel 统计可以用于检验二分因子变量和二分响应变量之间的条件独立性，条件是给定一个或多个分层（控制）变量定义的协变量模式。请注意：其他统计逐层计算，而 Cochran 和 Mantel-Haenszel 统计对所有层进行一次性计算。

交叉表：单元格显示

为帮助您发现数据中有助于显著性卡方检验的模式，交叉表过程显示期望频率和三种可测量观察的和期望的频率之间的差异的残差（偏差）。表的每个单元格可以包含选定计数、百分比值和残差的任意组合。

计数。如果行和列变量彼此独立，那么这是实际观察的个案数和期望的个案数。您可以选择隐藏小于指定整数的计数。隐藏的值将显示为 **<N**，其中 **N** 是指定的整数。指定的整数必须大于或等于 2，尽管允许指定值 0（表示不隐藏任何计数）。

比较列的比例。该选项将计算列属性的成对比较，并指出给定行中的哪对列明显不同。使用下标字母以 APA 样式格式在交叉表中标识显著性差异，并以 0.05 显著性水平对其进行计算。*注意*: 如果指定了此选项，但未选择观察计数或列百分比，那么观察计数将包括在交叉表表中，并通过 APA 样式的下标字母指示列比例检验结果。

- **调整 p 值（Bonferroni 方法）**。列比例的成对比较使用了 Bonferroni 修正，可在进行了多个比较后调整观察到的显著性水平。

百分比。百分比值可以跨行或沿列进行相加。还提供表（一层）中表示的个案总数的百分比值。*注意*: 如果在“计数”组中选中了**隐藏较小计数**，那么还将隐藏与隐藏计数相关联的百分比。

残差。未标准化的原始残差给出了观察值和期望值之间的差。还提供标准化残差和经过调整的标准化残差。

- *未标准化*。观察值与期望值之间的差。如果两个变量之间没有关系，那么期望值是期望在单元格中出现的个案数。如果行变量和列变量独立，那么正的残差表示单元格中的实际个案数多于期望的个案数。
- *标准化*。残差除以其标准差的估计。标准化残差也称为 Pearson 残差，它的平均值为 0，标准差为 1。
- *调节的标准化*。单元格的残差（观察值减去期望值）除以其标准误差的估计值。生成的标准化残差表示为平均值上下的标准差单位。

非整数权重。单元格计数通常为整数值，因为它们代表每个单元格中的个案个数。但是，如果数据文件当前按某个带小数值（例如 1.25）的权重变量进行加权，那么单元格计数也可能是小数值。在计算单元格计数之前可以进行截断或舍入，或为表显示和统计计算都使用小数单元格计数。

- *四舍五入单元格计数 (Round cell counts)*. 在计算任何统计之前，个案权重按原样使用，但单元格中的累积权重要四舍五入。
- *截断单元格计数 (Truncate cell counts)*. 在计算任何统计之前，个案权重按原样使用，但截断单元格中的累积权重。
- *四舍五入个案权重 (Round case weights)*. 在使用之前对个案权重进行四舍五入。
- *截断个案权重 (Truncate case weights)*. 在使用之前对个案权重进行截断。
- *无调节*。个案权重按原样使用且使用小数单元格计数。但是，当需要“精确”统计（仅由“精确检验”选项提供）时，在计算“精确”检验统计之前，单元格中的累积权重或者截断或者四舍五入。

交叉表: 格式

您可以按行变量值的升序或降序来排列行。

第 6 章 摘要

“摘要”过程为一个或多个分组变量的类别中的变量计算子组统计。所有级别的分组变量要进行交叉制表。您可以选择显示统计的顺序。还将显示跨所有类别的每个变量的汇总统计。每个类别中的数据值可以列出也可以不列出。对于大型数据集，可以选择只列出前 *n* 个个案。

示例。按地区和客户行业划分的平均产品销售额是多少？您可能会发现西部地区的平均销售额要略高于其他地区，西部地区的公司客户具有最高的平均销售额。

统计。合计、个案数、平均值、中位数、组内中位数、平均值的标准误差、最小值、最大值、范围、分组变量的第一个类别的变量值、分组变量的最后一个类别的变量值、标准差、方差、峰度、峰度标准误差、偏度、偏度标准误差、总和百分比、总个案数百分比、合计百分比、个案数百分比、几何平均值以及调和平均值。

摘要数据注意事项

数据。分组变量是分类变量，其值可以是数值或字符串。类别数应相当小。其他变量必须能排序。

假设。一些可选的子组统计（例如平均值和标准差）是基于正态理论的，适用于具有对称分布的定量变量。诸如中位数和范围之类的健壮性统计适用于定量变量，不管定量变量是否满足正态性假设。

获取个案摘要

1. 从菜单中选择:

   分析 > 报告 > 个案摘要...
2. 选择一个或多个变量。

根据需要，您可以:

- 选择一个或多个分组变量以将数据划分成子组。
- 单击**选项**以更改输出标题，在输出下面添加文字说明，或排除具有缺失值的个案。
- 单击**统计**获取可选的统计。
- 选择**显示个案**以列出每个子组中的个案。缺省情况下，系统只列出文件中的前 100 个个案。可以增大或减小“**将个案限制到前** *n* 个”中的 n 值，也可以取消选择该项以列出所有个案。

摘要选项

摘要允许您更改输出的标题或者添加文字说明，文字说明将显示在输出表下面。通过在文本中任何需要插入换行符的地方键入 \n，可以控制标题和文字说明中的换行。

您还可以选择显示或不显示总计的子标题，以及包含或排除在任何分析中使用的任何变量具有缺失值的个案。通常需要在输出中用句点或星号表示缺失个案。请输入要在值缺失的情况下显示的字符、短语或代码，否则不会对输出中的缺失个案应用特殊处理。

© Copyright IBM Corporation 1989, 2014

汇总统计

您可以为每个分组变量的每个类别中的变量选择下列一个或多个子组统计：合计、个案数、平均值、中位数、组内中位数、平均值的标准误差、最小值、最大值、范围、分组变量的第一个类别的变量值、分组变量的最后一个类别的变量值、标准差、方差、峰度、峰度标准误差、偏度、偏度标准误差、总和百分比、总个案数百分比、合计百分比、个案数百分比、几何平均值以及调和平均值。统计在“单元格统计”列表中的显示顺序就是它们将在输出中出现的顺序。还将显示跨所有类别的每个变量的汇总统计。

第一个 (First). 显示在数据文件中遇到的第一个数据值。

几何平均值 (Geometric Mean). 数据值的乘积的 n 次根，其中 n 代表个案数目。

组内中位数 (Grouped Median). 针对编码到组中的数据计算的中位数。例如，如果对于每个 30 年代的年龄数据的值都编码为 35，40 年代的编码为 45，依次类推，那么组内中位数是由已编码的数据计算得出的。

调和平均值 (Harmonic Mean). 在组中的样本大小不相等的情况下用来估计平均组大小。调和平均值是样本总数除以样本大小的倒数总和。

峰度 (Kurtosis). 观察值聚集在中点周围的程度的测量。对于正态分布，峰度统计的值为 0。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，leptokurtic 分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，platykurtic 分布的尾部比正态分布的尾部要薄。

最后一个 (Last). 显示在数据文件中遇到的最后一个数据值。

最大值 (Maximum). 数值变量的最大值。

平均值 (Mean). 集中趋势的测量。算术平均，总和除以个案个数。

中位数 (Median). 第 50 个百分位，大于该值和小于该值的个案数各占一半。如果个案个数为偶数，那么中位数是个案在以升序或降序排列的情况下最中间的两个个案的平均。中位数是集中趋势的测量，但对于远离中心的值不敏感（这与平均值不同，平均值容易受到少数多个非常大或非常小的值的影响）。

最小值 (Minimum). 数值变量的最小值。

N . 个案（观察值或记录）的数目。

总个案数的百分比。每个类别中的个案总数的百分比。

总和的百分比。每个类别中的总和百分比。

范围 (Range). 数值变量最大值和最小值之间的差；最大值减去最小值。

偏度 (Skewness). 分布的不对称性测量。正态分布是对称的，偏度值为 0。具有显著的正偏度的分布有很长的右尾。具有显著的负偏度的分布有很长的左尾。作为一个指导，当偏度值超过标准误差的两倍时，那么认为不具有对称性。

标准差 (Standard Deviation). 对围绕平均值的离差的测量。在正态分布中，68% 的个案在平均值的一倍标准差范围内，95% 的个案在平均值的两倍标准差范围内。例如，在正态分布中，如果平均年龄为 45，标准差为 10，那么 95% 的个案将处于 25 到 65 之间。

峰度标准误差 (Standard Error of Kurtosis). 峰度与其标准误差的比可用作正态性检验（即，如果比值小于 -2 或大于 +2，就可以拒绝正态性）。大的正峰度值表示分布的尾部比正态分布的尾部要长一些；负峰度值表示比较短的尾部（变为像框状的均匀分布尾部）。

平均值的标准误差 (Standard Error of Mean). 取自同一分布的样本与样本之间的平均值之差的测量。它可以用来粗略地将观察到的平均值与假设值进行比较（即，如果差与标准误差的比值小于 -2 或大于 +2，那么可以断定两个值不同）。

偏度标准误差 (Standard Error of Skewness). 偏度与其标准误差的比可用作正态性检验（即，如果比值小于 -2 或大于 +2，就可以拒绝正态性）。大的正偏度值表示长右尾；极负值表示长左尾。

总和 (Sum). 所有带有非缺失值的个案的值的合计或总计。

方差 (Variance). 对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。度量方差的单位是变量本身的单位的平方。

第 7 章 平均值

平均值过程计算一个或多个自变量类别中因变量的子组平均值和相关的单变量统计。您也可以获得单向方差检验分析、eta 和线性相关度相关检验。

示例。度量三类不同的烹调油所吸收的平均脂肪量，并执行单向方差检验分析，查看平均值是否不同。

统计。合计、个案数、平均值、中位数、组内中位数、平均值的标准误差、最小值、最大值、范围、分组变量的第一个类别的变量值、分组变量的最后一个类别的变量值、标准差、方差、峰度、峰度标准误差、偏度、偏度标准误差、总和百分比、总个案数百分比、合计百分比、个案数百分比、几何平均值以及调和平均值。选项包括方差分析、eta、eta 平方和线性相关度 R 和 R^2 检验。

平均值数据注意事项

数据。因变量为定量变量，自变量为分类变量。分类变量的值可以为数字，也可以为字符串。

假设。一些可选的子组统计（例如平均值和标准差）是基于正态理论的，适用于具有对称分布的定量变量。稳健统计（如中位数）适用于可能符合或可能不符合正态假设的定量变量。方差分析对于偏离正态是稳健的，但每个单元格中的数据应该是对称的。方差分析还假定各组来自方差相同的总体。要检验这种假定，请使用 Levene 的方差同质性检验，可以从单因素方差分析过程中获得。

获得子组平均值

1. 从菜单中选择:

   分析 > 比较平均值 > 平均值...
2. 选择一个或多个因变量。
3. 使用下列一种方法选择分类自变量:
   - 选择一个或多个自变量。显示每个自变量的单独的结果。
   - 选择一层或多层自变量。每一层都将进一步细分样本。如果在层 1 中有一个自变量，层 2 中也有一个自变量，结果就显示为一个交叉的表，而不是对每个自变量显示一个独立的表。
4. 或者，单击**选项**选择可选统计、方差表的分析、eta、eta 平方、R 和 R^2。

平均值: 选项

您可以为每个分组变量的每个类别中的变量选择下列一个或多个子组统计: 合计、个案数、平均值、中位数、组内中位数、平均值的标准误差、最小值、最大值、范围、分组变量的第一个类别的变量值、分组变量的最后一个类别的变量值、标准差、方差、峰度、峰度标准误差、偏度、偏度标准误差、总和百分比、总个案数百分比、合计百分比、个案数百分比、几何平均值以及调和平均值。您可更改子组统计出现的顺序。统计在“单元格统计”列表中出现的顺序是它们在输出中显示的顺序。还将显示跨所有类别的每个变量的汇总统计。

第一个 (First). 显示在数据文件中遇到的第一个数据值。

几何平均值 (Geometric Mean). 数据值的乘积的 n 次根，其中 n 代表个案数目。

组内中位数 (Grouped Median). 针对编码到组中的数据计算的中位数。例如，如果对于每个 30 年代的年龄数据的值都编码为 35，40 年代的编码为 45，依次类推，那么组内中位数是由已编码的数据计算得出的。

© Copyright IBM Corporation 1989, 2014

调和平均值 (Harmonic Mean). 在组中的样本大小不相等的情况下用来估计平均组大小。调和平均值是样本总数除以样本大小的倒数总和。

峰度 (Kurtosis). 观察值聚集在中点周围的程度的测量。对于正态分布，峰度统计的值为 0。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，leptokurtic 分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，platykurtic 分布的尾部比正态分布的尾部要薄。

最后一个 (Last). 显示在数据文件中遇到的最后一个数据值。

最大值 (Maximum). 数值变量的最大值。

平均值 (Mean). 集中趋势的测量。算术平均，总和除以个案个数。

中位数 (Median). 第 50 个百分位，大于该值和小于该值的个案数各占一半。如果个案个数为偶数，那么中位数是个案在以升序或降序排列的情况下最中间的两个个案的平均。中位数是集中趋势的测量，但对于远离中心的值不敏感（这与平均值不同，平均值容易受到少数多个非常大或非常小的值的影响）。

最小值 (Minimum). 数值变量的最小值。

N . 个案（观察值或记录）的数目。

合计 N % (Percent of total N). 每个类别中的个案总数的百分比。

总和的百分比 (Percent of total sum). 每个类别中的总和百分比。

范围 (Range). 数值变量最大值和最小值之间的差；最大值减去最小值。

偏度 (Skewness). 分布的不对称性测量。正态分布是对称的，偏度值为 0。具有显著的正偏度的分布有很长的右尾。具有显著的负偏度的分布有很长的左尾。作为一个指导，当偏度值超过标准误差的两倍时，那么认为不具有对称性。

标准差 (Standard Deviation). 对围绕平均值的离差的测量。在正态分布中，68% 的个案在平均值的一倍标准差范围内，95% 的个案在平均值的两倍标准差范围内。例如，在正态分布中，如果平均年龄为 45，标准差为 10，那么 95% 的个案将处于 25 到 65 之间。

峰度标准误差 (Standard Error of Kurtosis). 峰度与其标准误差的比可用作正态性检验（即，如果比值小于 -2 或大于 +2，就可以拒绝正态性）。大的正峰度值表示分布的尾部比正态分布的尾部要长一些；负峰度值表示比较短的尾部（变为像框状的均匀分布尾部）。

平均值的标准误差 (Standard Error of Mean). 取自同一分布的样本与样本之间的平均值之差的测量。它可以用来粗略地将观察到的平均值与假设值进行比较（即，如果差与标准误差的比值小于 -2 或大于 +2，那么可以断定两个值不同）。

偏度标准误差 (Standard Error of Skewness). 偏度与其标准误差的比可用作正态性检验（即，如果比值小于 -2 或大于 +2，就可以拒绝正态性）。大的正偏度值表示长右尾；极负值表示长左尾。

总和 (Sum). 所有带有非缺失值的个案的值的合计或总计。

方差 (Variance). 对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。度量方差的单位是变量本身的单位的平方。

第一层的统计

Anova 表和 eta。显示单因素方差分析表，并为第一层中的每个自变量计算 eta 和 eta 平方（相关性测量）。

线性相关度检验。计算与线性和非线性成分相关联的平方和、自由度和均方，以及 F 比、R 和 R 方。如果自变量为短字符串，那么不计算线性相关度。

第 8 章 OLAP 多维数据集

OLAP（联机分析处理）多维数据集过程计算一个或多个分类分组变量类别中连续摘要变量的总和、平均值和其他单变量统计。在表中为每个分组变量的每个类别创建单独的层。

示例。不同区域的总销售额和平均销售额以及区域内的产品线。

统计。和、个案数、平均值、中位数、组内中位数、平均值的标准误差、最小值、最大值、范围、分组变量的第一类别的变量值、分组变量的最后一个类别的变量值、标准差、方差、峰度、峰度标准误差、偏度、偏度标准误差、个案总数的百分比、总和的百分比、分组变量中个案总数的百分比、分组变量中总和的百分比、几何平均值和调和平均值。

OLAP 多维数据集数据注意事项

数据。摘要变量为定量变量（定距或者定比度量的连续变量），分组变量为分类变量。分类变量的值可以为数字，也可以为字符串。

假设。一些可选的子组统计（例如平均值和标准差）是基于正态理论的，适用于具有对称分布的定量变量。稳健统计（如中位数和范围）适用于可能符合或可能不符合正态假设的定量变量。

获得 OLAP 多维数据集

1. 从菜单中选择:

   分析 > 报告 > OLAP 多维数据集...
2. 选择一个或多个连续摘要变量。
3. 选择一个或多个分类分组变量。

或者:

- 选择不同的汇总统计（单击**统计**）。在选择汇总统计之前，必须选择一个或多个分组变量。
- 计算变量对和由分组变量定义的组对之间的差（单击**差分**）。
- 创建定制表标题（单击**标题**）。
- 隐藏小于指定整数的计数。隐藏的值将显示为 **<N**，其中 **N** 是指定的整数。指定的整数必须大于或等于 2。

OLAP 多维数据集: 统计

您可以为每个分组变量的每个类别中的摘要变量选择下列一个或多个子组统计: 合计、个案数、平均值、中位数、组内中位数、平均值的标准误差、最小值、最大值、范围、分组变量的第一个类别的变量值、分组变量的最后一个类别的变量值、标准差、方差、峰度、峰度标准误差、偏度、偏度标准误差、总个案数百分比、总和百分比、分组变量中的总个案数百分比、分组变量中的总和百分比、几何平均值以及调和平均值。

您可更改子组统计出现的顺序。统计在“单元格统计”列表中出现的顺序是它们在输出中显示的顺序。还将显示跨所有类别的每个变量的汇总统计。

第一个 (First). 显示在数据文件中遇到的第一个数据值。

几何平均值 (Geometric Mean). 数据值的乘积的 n 次根，其中 n 代表个案数目。

© Copyright IBM Corporation 1989, 2014

组内中位数 (Grouped Median). 针对编码到组中的数据计算的中位数。例如，如果对于每个 30 年代的年龄数据的值都编码为 35，40 年代的编码为 45，依次类推，那么组内中位数是由已编码的数据计算得出的。

调和平均值 (Harmonic Mean). 在组中的样本大小不相等的情况下用来估计平均组大小。调和平均值是样本总数除以样本大小的倒数总和。

峰度 (Kurtosis). 观察值聚集在中点周围的程度的测量。对于正态分布，峰度统计的值为 0。正峰度值表示相对于正态分布，观察值在分布中心的聚集更多，同时尾部更薄，直到分布极值。在这一点，leptokurtic 分布的尾部比正态分布的尾部要厚。负峰度值表示相对于正态分布，观察值聚集得少并且尾部较厚，直到分布极值。在这一点，platykurtic 分布的尾部比正态分布的尾部要薄。

最后一个 (Last). 显示在数据文件中遇到的最后一个数据值。

最大值 (Maximum). 数值变量的最大值。

平均值 (Mean). 集中趋势的测量。算术平均，总和除以个案个数。

中位数 (Median). 第 50 个百分位，大于该值和小于该值的个案数各占一半。如果个案个数为偶数，那么中位数是个案在以升序或降序排列的情况下最中间的两个个案的平均。中位数是集中趋势的测量，但对于远离中心的值不敏感（这与平均值不同，平均值容易受到少数多个非常大或非常小的值的影响）。

最小值 (Minimum). 数值变量的最小值。

N . 个案（观察值或记录）的数目。

数量的百分比 (Percent of N in). 其他分组变量的类别内指定分组变量的个案数的百分比。如果只有一个分组变量，那么此值与总个案数百分比相同。

合计百分比. 其他分组变量的类别内指定分组变量的总和的百分比。如果只有一个分组变量，那么此值与总和百分比相同。

总个案数的百分比. 每个类别中的个案总数的百分比。

总和的百分比. 每个类别中的总和百分比。

范围 (Range). 数值变量最大值和最小值之间的差；最大值减去最小值。

偏度 (Skewness). 分布的不对称性测量。正态分布是对称的，偏度值为 0。具有显著的正偏度的分布有很长的右尾。具有显著的负偏度的分布有很长的左尾。作为一个指导，当偏度值超过标准误差的两倍时，那么认为不具有对称性。

标准差 (Standard Deviation). 对围绕平均值的离差的测量。在正态分布中，68% 的个案在平均值的一倍标准差范围内，95% 的个案在平均值的两倍标准差范围内。例如，在正态分布中，如果平均年龄为 45，标准差为 10，那么 95% 的个案将处于 25 到 65 之间。

峰度标准误差 (Standard Error of Kurtosis). 峰度与其标准误差的比可用作正态性检验（即，如果比值小于 -2 或大于 +2，就可以拒绝正态性）。大的正峰度值表示分布的尾部比正态分布的尾部要长一些；负峰度值表示比较短的尾部（变为像框状的均匀分布尾部）。

平均值的标准误差 (Standard Error of Mean). 取自同一分布的样本与样本之间的平均值之差的测量。它可以用来粗略地将观察到的平均值与假设值进行比较（即，如果差与标准误差的比值小于 -2 或大于 +2，那么可以断定两个值不同）。

偏度标准误差 (Standard Error of Skewness). 偏度与其标准误差的比可用作正态性检验（即，如果比值小于 -2 或大于 +2，就可以拒绝正态性）。大的正偏度值表示长右尾；极负值表示长左尾。

总和 (Sum). 所有带有非缺失值的个案的值的合计或总计。

方差 (Variance). 对围绕平均值的离差的测量，值等于与平均值的差的平方和除以个案数减一。度量方差的单位是变量本身的单位的平方。

OLAP 多维数据集差

该对话框允许您计算摘要变量间或由分组变量定义的组间的百分比和算术差。将计算“OLAP 多维数据集: 统计”对话框中选定的所有度量的差。

变量之间移动的差值。计算变量对之间的差值。每一对中第一个变量的汇总统计值减去第二个变量（减去的变量）的汇总统计值。对百分比差而言，减去的变量的摘要变量值用作分母。在可以指定变量间的差之前，您必须在主对话框中至少选择两个摘要变量。

个案组之间的差值。计算由分组变量定义的组对间的差。每一对中第一个类别的汇总统计值减去第二个类别（减去的类别）的汇总统计值。百分比差将减去的类别的汇总统计值作为分母。必须在主对话框中选择一个或多个分组变量，之后才能指定组之间的差分。

OLAP 多维数据集: 标题

可更改输出标题或添加在输出表下面显示的文字说明。在文本中，您还可在想插入换行符的地方输入 `\n` 以控制标题和文字说明的换行。

第 9 章 t 检验

t 检验

有三类 *t* 检验可用:

独立样本 t 检验（双样本 t 检验）。比较两组个案中一个变量的平均值。提供了每组的描述统计和 Levene 方差等同性检验，以及相等和不等方差 *t* 值和平均值差分的 95% 置信区间。

配对样本 t 检验（相关 t 检验）。比较单个组的两个变量的平均值。此检验还用于匹配对或个案控制研究设计。输出包括检验变量的描述统计、变量之间的相关性、配对差分的描述统计、*t* 检验和 95% 置信区间。

单样本 t 检验。将一个变量的平均值与已知值或假设值进行比较。检验变量的描述统计随 *t* 检验一起显示。检验变量的平均值和假设的检验值之间差的 95% 置信区间是缺省输出的一部分。

独立样本 T 检验

“独立样本 T 检验”过程比较两组个案的平均值。理想的情况下，对于此检验，主体应随机地分配到两个组中，以便响应的任何差别是由于处理（或缺少处理）而非其他因素造成的。而比较男性和女性的平均收入则不属于此情况。人不是随机指定为男性或女性的。在这些情况下，您应确保其他因素中的差别没有掩饰或夸大平均值中的显著性差异。平均收入的差值还可能受诸如教育之类的因素影响（而非仅仅受性别影响）。

示例。高血压病人随机地分配到安慰剂组和治疗组。安慰剂主体主要接受无效的药丸，而治疗主体主要接受一种期望能降低血压的新药。在主体经过两个月的治疗之后，使用双样本 *t* 检验比较安慰剂组和治疗组的平均血压。每名病人测量一次并归属于一个组。

统计。对于每个变量: 样本大小、平均值、标准差以及平均值的标准误差。对于平均值的差异: 平均值、标准误差和置信区间（您可以指定置信度级别）。检验: Levene 方差相等性检验以及平均值相等性的汇聚方差和分离方差 *t* 检验。

独立样本 T 检验数据注意事项

数据。感兴趣的定量变量的值位于数据文件的单独一列中。此过程使用具有两个值的分组变量将个案分成两个组。分组变量可以是数值（诸如 1 和 2，或者 6.25 和 12.5 之类的值），也可以是短字符串（例如 *yes* 和 *no*）。作为备选方法，您可以使用定量变量（例如*年龄*）来将个案分成两个组，方法是指定一个分割点（分割点 21 将*年龄*分成 21 岁以下组和 21 岁及以上组）。

假设。对于相等方差 *t* 检验，观察值应是来自具有相等的总体方差的正态分布的独立随机样本。对于不等方差 *t* 检验，观察值应是来自正态分布的独立随机样本。双样本 *t* 检验对于偏离正态性是相当稳健的。当以图形方式检查分布时，请检查以确保它们对称且没有离群值。

获取独立样本 T 检验

1. 从菜单中选择:

   分析 > 比较平均值 > 独立样本 T 检验...

2. 选择一个或多个定量检验变量。为每个变量单独计算 *t* 检验。
3. 选择单个分组变量，然后单击**定义组**为想要比较的组指定两个代码。

© Copyright IBM Corporation 1989, 2014

4. 或者，单击**选项**以控制缺失数据的处理和置信区间的置信度。

独立样本 T 检验：定义组

对于数值分组变量，通过指定两个值或一个分割点为 *t* 检验定义两个组：

- **使用指定值。**为“组 1”输入一个值，为“组 2”输入另一个值。具有任何其他值的个案将从分析中排除。数字不需要是整数（例如 6.25 和 12.5 也有效）。
- **分割点**。输入一个将分组变量的值分成两组的数字。值小于分割点的所有个案组成一个组，值大于等于分割点的个案组成另一个组。

对于字符串分组变量，请为“组 1”输入一个字符串，并为“组 2”输入另一个值，例如 *yes* 和 *no*。具有其他字符串的个案将从分析中排除。

独立样本 T 检验：选项

置信区间。缺省情况下，显示平均值中的差的 95% 置信区间。可输入 1 到 99 之间的值以请求不同的置信度。

缺失值。当您检验多个变量，并且一个或多个变量的数据缺失时，您可以指示过程包含（或排除）哪些个案。

- **按分析排除个案**。每个 *t* 检验均使用对于检验的变量具有有效数据的全部个案。样本大小可能随检验的不同而不同。
- **按列表排除个案**。每个 *t* 检验只使用对于在请求的 *t* 检验中使用的所有变量具有有效数据的个案。样本大小在各个检验之间恒定。

配对样本 T 检验

“配对样本 T 检验”过程比较单独一组的两个变量的平均值。此过程计算每个个案的两个变量的值之间的差值，并检验平均差值是否非 0。

示例。在对高血压的研究中，在研究开始测量所有病人的血压，在治疗之后再次测量血压。这样，每个主体有两个测量值，它们通常称为*之前* 测量值和*之后* 测量值。使用该检验的另一个设计是匹配对或个案控制研究，在此研究中，数据文件中的每个记录包含病人的反应，以及该病人的匹配控制主体的反应。在血压研究中，病人和控制可以按年龄匹配（75 岁的病人与 75 岁的控制组成员匹配）。

统计。对于每个变量：平均值、样本大小、标准差以及平均值的标准误差。对于每一对变量：相关性、平均值的平均差、*t* 检验以及平均值差值的置信区间（您可以指定置信度级别）。平均值差值的标准差和标准误差。

配对样本 T 检验数据注意事项

数据。对于每个配对检验，指定两个定量变量（定距测量级别或定比测量级别）。对于匹配对或个案控制研究，每个检验主体的响应及其匹配的控制主体的响应必须在数据文件的相同个案中。

假设。每对的观察值应在相同的条件下得到。平均值差值应是正态分布的。每个变量的方差可以相等也可以不等。

获取配对样本 T 检验

1. 从菜单中选择：

   分析 > 比较平均值 > 配对样本 T 检验...

2. 选择一对或多对变量。
3. 或者，单击**选项**以控制缺失数据的处理和置信区间的置信度。

配对样本 T 检验: 选项

置信区间。缺省情况下，显示平均值中的差的 95% 置信区间。可输入 1 到 99 之间的值以请求不同的置信度。

缺失值。当您检验多个变量，并且一个或多个变量的数据缺失时，您可以指示过程包含（或排除）哪些个案:

- **按分析排除个案**。每个 *t* 检验均使用对于检验的变量对具有有效数据的全部个案。样本大小可能随检验的不同而不同。
- **按列表排除个案**。每个 *t* 检验只使用对于所有检验的变量对具有有效数据的个案。样本大小在各个检验之间恒定。

T-TEST 命令的附加功能

使用命令语法语言还可以:

- 通过运行单个命令同时生成单样本 t 检验和独立样本 t 检验。
- 在配对 t 检验中，针对列表中的每个变量检验一个变量（使用 PAIRS 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

单样本 T 检验

“单样本 T 检验”过程检验单个变量的平均值是否与指定的常数不同。

示例。某研究人员可能想要检验一组学生的平均 IQ 得分是否不等于 100。或者，某谷物制造商可以从生产线取箱子的样本，并检查样本的平均重量在 95% 的置信度下是否不等于 1.3 磅。

统计。对于每个检验变量: 平均值、标准差以及平均值的标准误差。每个数据值和假设的检验值之间的平均差、检验此差为 0 的 *t* 检验、以及此差的置信区间（您可以指定置信度）。

单样本 T 检验数据注意事项

数据。要对照假设的检验值检验定量变量的值，可选择一个定量变量并输入一个假设的检验值。

假设。此检验假设数据正态分布；但是，此检验对偏离正态性是相当稳健的。

获取单样本 T 检验

1. 从菜单中选择:

   分析 > 比较平均值 > 单样本 T 检验...
2. 选择要对照同一假设值进行检验的一个或多个变量。
3. 输入一个数值检验值，每个样本平均值要与之进行比较。
4. 或者，单击**选项**以控制缺失数据的处理和置信区间的置信度。

单样本 T 检验: 选项

置信区间。缺省情况下，显示平均值和假设的检验值之差的 95% 置信区间。可输入 1 到 99 之间的值以请求不同的置信度。

缺失值。当您检验多个变量，并且一个或多个变量的数据缺失时，您可以指示过程包含（或排除）哪些个案。

- **按分析排除个案**。每个 *t* 检验均使用对于检验的变量具有有效数据的全部个案。样本大小可能随检验的不同而不同。

- **按列表排除个案**。每个 *t* 检验只使用对于在任何请求的 *t* 检验中使用的所有变量都具有有效数据的个案。样本大小在各个检验之间恒定。

T-TEST 命令的附加功能

使用命令语法语言还可以:

- 通过运行单个命令同时生成单样本 t 检验和独立样本 t 检验。
- 在配对 t 检验中，针对列表中的每个变量检验一个变量（使用 PAIRS 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

T-TEST 命令的附加功能

使用命令语法语言还可以:

- 通过运行单个命令同时生成单样本 t 检验和独立样本 t 检验。
- 在配对 t 检验中，针对列表中的每个变量检验一个变量（使用 PAIRS 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 10 章 单因素 ANOVA

“单因素 ANOVA”过程按照单因子变量（自变量）生成对定量因变量的单向方差检验分析。方差分析用于检验数个平均值相等的假设。这种方法是双样本 *t* 检验的扩展。

除了确定平均值间存在着差值外，您可能还想知道哪些平均值之间存在着差值。比较平均值有两类检验方法：先验对比和事后检验。对比是在试验开始*前*进行的检验，而事后检验则是在试验结束*后*进行的。您也可以检验各个类别的趋势。

示例。炸面包圈在烹制过程中吸收的脂肪量各不相同。我们设计了一个涉及三种脂肪的实验：花生油、玉米油和猪油。花生油和玉米油是不饱和脂肪，而猪油是饱和脂肪。除了确定吸收的脂肪量是否因使用的脂肪类型而异外，您还可以建立一个先验对比，确定吸收的脂肪量是否也因饱和脂肪和不饱和脂肪而异。

统计。对于每个组：个案数、平均值、标准差、平均值的标准误差、最小值、最大值以及平均值的 95% 置信区间。Levene 方差同质性检验、每个因变量的方差分析表和平均值相等性稳健测试、用户指定的先验对比以及事后范围检验和多重比较：Bonferroni、Sidak、Tukey 真实显著性差异、Hochberg GT2、Gabriel、Dunnett、Ryan-Einot-Gabriel-Welsch *F* 检验 (R-E-G-W *F*)、Ryan-Einot-Gabriel-Welsch 范围检验 (R-E-G-W *Q*)、Tamhane T2、Dunnett T3、Games-Howell、Dunnett *C*、Duncan 多范围检验、Student-Newman-Keuls (S-N-K)、Tukey *b*、Waller-Duncan、Scheffé 和最小显著性差异。

单因素 ANOVA 数据注意事项

数据。因子变量值应为整数，而因变量应为定量变量（区间测量级别）。

假设。每个组是来自正态总体的独立随机样本。尽管数据应对称，但方差分析对于偏离正态性是稳健的。各组应来自方差相等的总体。为了检验这种假设，请使用 Levene 的方差同质性检验。

获取单向方差检验分析

1. 从菜单中选择:

   分析 > 比较平均值 > 单因素 ANOVA...
2. 选择一个或多个因变量。
3. 选择一个自变量因子变量。

单因素 ANOVA: 对比

您可以将组间平方和划分成趋势成分，或者指定先验对比。

多项式。将组间平方和划分成趋势成分。可以检验因变量在因子变量的各顺序水平间的趋势。例如，您可以检验各个顺序级别的最高工资水平间的线性趋势（上升或下降）。

- **度**。可以选择 1 度、2 度、3 度、4 度或 5 度多项式。

系数。用户指定的用 *t* 统计检验的先验对比。为因子变量的每个组（类别）输入一个系数，每次输入后单击**添加**。每个新值都添加到系数列表的底部。要指定其他对比组，请单击**下一个**。用**下一个**和**上一个**在各组对比间移动。

© Copyright IBM Corporation 1989, 2014

系数的顺序很重要，因为该顺序与因子变量的类别值的升序相对应。列表中的第一个系数与因子变量的最低组值相对应，而最后一个系数与最高值相对应。例如，如果有 6 类因子变量，那么系数 -1、0、0、0、0.5 和 0.5 将第一组与第五和第六组进行对比。对于大多数应用程序而言，各系数的和应为 0。系数和不是 0 的集也可以使用，但是会出现一条警告消息。

单因素 ANOVA: 事后检验

一旦确定平均值间存在差值，两两范围检验和成对多重比较就可以确定哪些平均值存在差值了。范围检验识别彼此间没有差值的同类平均值子集。成对多重比较检验每一对平均值之间的差分，并得出一个矩阵，其中星号指示在 0.05 的 alpha 水平上的组平均值明显不同。

假定方差齐性

Tukey 真实显著性差异检验、Hochberg GT2、Gabriel 和 Scheffé 是多重比较检验和范围检验。其他可用的范围检验为 Tukey 的 *b*、S-N-K (Student-Newman-Keuls)、Duncan、R-E-G-W *F*（Ryan-Einot-Gabriel-Welsch *F* 检验）、R-E-G-W *Q*（Ryan-Einot-Gabriel-Welsch 范围检验）和 Waller-Duncan。可用的多重比较检验包括 Bonferroni、Tukey 真实显著性差异检验、Sidak、Gabriel、Hochberg、Dunnett、Scheffé 和 LSD（最小显著性差异）。

- *LSD*。使用 t 检验执行组平均值之间的所有成对比较。对多个比较的误差率不做调整。
- *Bonferroni*. 使用 t 检验在组平均值之间执行成对比较，但通过将每次检验的误差率设置为实验性质的误差率除以检验总数来控制总体误差率。这样，根据进行多个比较的实情对观察的显著性水平进行调整。
- *Sidak*。 基于 t 统计的成对多重比较检验。Sidak 调整多重比较的显著性水平，并提供比 Bonferroni 更严密的边界。
- *Scheffe*. 为平均值的所有可能的成对组合执行并发的联合成对比较。使用 F 取样分布。可用来检查组平均值的所有可能的线性组合，而非仅限于成对组合。
- *R-E-G-W F*。基于 F 检验的 Ryan-Einot-Gabriel-Welsch 多步进过程。
- *R-E-G-W Q*。基于 Student 化的范围的 Ryan-Einot-Gabriel-Welsch 多步进过程。
- *S-N-K*。使用 Student 化的范围分布在平均值之间进行所有成对比较。它还使用步进式过程比较具有相同样本大小的同类子集内的平均值对。平均值按从高到低排序，首先检验极端差分。
- *Tukey*。使用 Student 化的范围统计量进行组间所有成对比较。将试验误差率设置为所有成对比较的集合的误差率。
- *Tukey b*。使用 Student 化的范围分布在组之间进行成对比较。临界值是 Tukey's 真实显著性差异检验的对应值与 Student-Newman-Keuls 的平均数。
- *Duncan*。使用与 Student-Newman-Keuls 检验所使用的完全一样的逐步顺序成对比较，但要为检验的集合的误差率设置保护水平，而不是为单个检验的误差率设置保护水平。使用 Student 化的范围统计量。
- *Hochberg GT2*。使用学生化最大模数的多重比较和范围检验。与 Tukey's 真实显著性差异检验相似。
- *Gabriel*。使用学生化最大模数的成对比较检验，并且当单元格大小不相等时，它通常比 Hochberg's GT2 更为强大。当单元格大小变化过大时，Gabriel 检验可能会变得随意。
- *Waller-Duncan*。基于 t 统计的多比较检验；使用 Bayesian 方法。
- *Dunnett*。成对多重比较 t 检验，它对照单个控制平均值来比较处理集合。最后一个类别是缺省的控制类别。另外，您还可以选择第一个类别。**双面**检验任何水平（除了控制类别外）的因子的平均值是否不等于控制类别的平均值。**<控制**检验任何水平的因子的平均值是否小于控制类别的平均值。**> 控制**检验任何水平的因子的平均值是否大于控制类别的平均值。

未假定方差齐性

不假设方差相等的多重比较检验有 Tamhane 的 T2、Dunnett T3、Games-Howell 和 Dunnett 的 *C*。

- *Tamhane T2*。基于 t 检验的保守成对比较。当方差不相等时，适合使用此检验。
- *Dunnett T3*。基于学生化最大值模数的成对比较检验。当方差不相等时，适合使用此检验。
- *Games-Howell*。有时会变得随意的成对比较检验。当方差不相等时，适合使用此检验。
- *Dunnett C*。基于 Student 化的范围的成对比较检验。当方差不相等时，适合使用此检验。

注: 如果取消选择“表格属性”对话框（在激活的透视表中，从“格式”菜单选择**表格属性**）中的**隐藏空行列**，可能更容易看懂事后检验的输出。

单因素 ANOVA: 选项

统计。选择下列各项的一个或多个:

- **描述性**。计算每组中每个因变量的个案数、平均值、标准差、平均值的标准误差、最小值、最大值和 95% 置信区间。
- **固定和随机效果**。显示固定效应模型的标准差、标准误差和 95% 置信区间，以及随机效应模型的标准误差、95% 置信区间和成分间方差估计。
- **方差同质性检验**。计算 Levene 统计以检验组方差是否相等。该检验独立于正态的假设。
- **Brown-Forsythe**。计算 Brown-Forsythe 统计以检验组平均值是否相等。当方差相等的假设不成立时，这种统计优于 *F* 统计。
- **Welch**。计算 Welch 统计以检验组平均值是否相等。当方差相等的假设不成立时，这种统计优于 *F* 统计。

平均值图。显示一个绘制子组平均值的图表（每组的平均值由因子变量的值定义）。

缺失值。控制对缺失值的处理。

- **按分析排除个案**。给定分析中的因变量或因子变量有缺失值的个案不用于该分析。而且，也不使用超出为因子变量指定的范围的个案。
- **按列表排除个案**。因子变量有缺失值的个案，或包括在主对话框中的因变量列表上的任何因变量的值缺失的个案都排除在所有分析之外。如果尚未指定多个因变量，那么这个选项不起作用。

ONEWAY 命令的附加功能

使用命令语法语言还可以:

- 获取固定效应和随机效应统计。固定效应模型的标准差、平均值标准误差和 95% 置信区间。随机效应模型的标准误差、95% 置信区间和成分间方差估计（使用 `STATISTICS=EFFECTS`）。
- 指定最小显著性差异的 Alpha 水平、Bonferroni、Duncan 和 Scheffé 多重比较检验（使用 `RANGES` 子命令）。
- 写平均值矩阵、标准差和频率，或者读平均值矩阵、频率、汇聚方差和汇聚方差的自由度。可以使用这些矩阵代替原始数据，以获取单向方差检验分析（使用 `MATRIX` 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 11 章 GLM 单变量分析

“GLM 单变量”过程通过一个或多个因子和/或变量，为一个因变量提供回归分析和方差分析。因子变量将总体划分成组。通过使用此“一般线性模型”过程，您可以检验关于其他变量对单个因变量的各个分组的平均值的效应的原假设。您可以调查因子之间的交互以及个别因子的效应，它们之中有些可能是随机的。另外，还可以包含协变量的效应以及协变量与因子的交互。对于回归分析，自变量（预测变量）指定为协变量。

平衡与非平衡模型均可进行检验。如果模型中的每个单元格包含相同的个案数，那么设计是平衡的。除了检验假设，“GLM 单变量”还生成参数估计值。

常用的先验对比可用于执行假设检验。另外，在整体的 *F* 检验已显示显著性之后，可以使用事后检验评估指定平均值之间的差值。估计边际平均值为模型中的单元格给出了预测平均值的估计值，且这些平均值的概要图（交互图）允许您容易地对其中一些关系进行可视化。

残差、预测值、Cook 距离以及杠杆值可以另存为数据文件中检查假设的新变量。

WLS 权重允许您指定用来为加权最小二乘 (WLS) 分析赋予观察值不同权重的变量，这样也许可以补偿不同的测量精度。

示例。数年来一直收集芝加哥马拉松赛中各个选手的数据。每名选手到达终点的时间是因变量。其他因子包括天气（冷、舒适或热）、训练月数、以前参加马拉松赛的次数以及性别。年龄被视为协变量。您可能会发现性别是一个显著作用，性别与天气的交互也是显著的。

方法。类型 I、类型 II、类型 III 和类型 IV 的平方和可用来评估不同的假设。类型 III 是缺省值。

统计。事后范围检验和多重比较: 最小显著性差异、Bonferroni、Sidak、Scheffé、Ryan-Einot-Gabriel-Welsch 多重 *F*、Ryan-Einot-Gabriel-Welsch 多范围、Student-Newman-Keuls、Tukey 真实显著性差异、Tukey *b*、Duncan、Hochberg GT2、Gabriel、Waller-Duncan *t* 检验、Dunnett（单侧和双侧）、Tamhane T2、Dunnett T3、Games-Howell 和 Dunnett *C*。描述统计: 所有单元格中的所有因变量的观察平均值、标准差和计数。Levene 的方差同质性检验。

图。分布-水平图、残差图以及概要图（交互）。

GLM 单变量数据注意事项

数据。因变量是定量变量。因子是分类变量。它们可以具有数字值或最多 8 个字符的字符串值。协变量是与因变量相关的定量变量。

假设。数据是来自正态总体的随机样本，在总体中，所有单元格方差相同。尽管数据应对称，但方差分析对于偏离正态性是稳健的。要检查假设，您可以使用方差同质性检验和分布-水平图。您还可以检查残差和残差图。

获取 GLM 单变量表

1. 从菜单中选择:

   分析 > 一般线性模型 > 单变量...

2. 选择一个因变量。
3. 为“固定因子”、“随机因子”和“协变量”选择变量（如果适用于您的数据）。

© Copyright IBM Corporation 1989, 2014

4. 可选地，您还可以使用“WLS 权重”为加权最小二乘分析指定权重变量。如果加权变量的值为 0、负数或缺失，那么将该个案从分析中排除。已用在模型中的变量不能用作加权变量。

GLM 模型

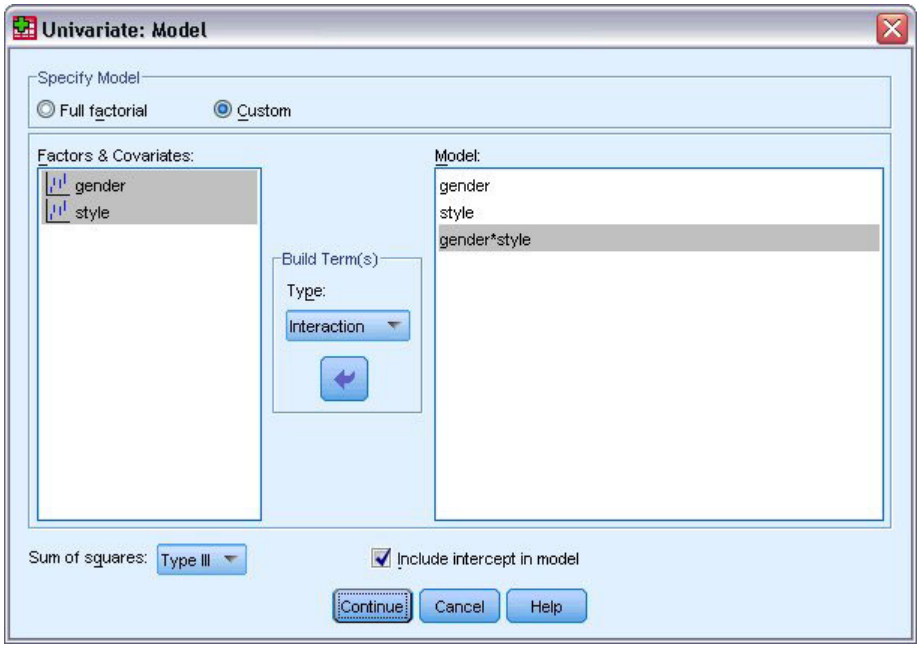

图 1. “单变量模型”对话框

指定模型。全因子模型包含所有因子主效应、所有协变量主效应以及所有因子间交互。它不包含协变量交互。选择**定制**可以仅指定其中一部分的交互或指定因子协变量交互。必须指定要包含在模型中的所有项。

因子与协变量。列出因子与协变量。

模型。模型取决于数据的性质。选择**定制**之后，您可以选择分析中感兴趣的主效应和交互效应。

平方和。计算平方和的方法。对于没有缺失单元格的平衡或非平衡模型，类型 III 平方和法最常用。

在模型中包含截距。模型中通常包含截距。如果您可以假设数据穿过原点，那么可以排除截距。

建立项

对于选定因子和协变量:

交互。创建所有选定变量的最高级交互项。这是缺省值。

主效应。为每个选定的变量创建主效应项。

所有二阶。创建选定变量的所有可能的双向交互。

所有三阶。创建选定变量的所有可能的三阶交互。

所有四阶。创建选定变量的所有可能的四阶交互。

所有五阶。创建选定变量的所有可能的五阶交互。

平方和

对于该模型，您可以选择平方和类型。类型 III 最常用，并且是缺省类型。

类型 I。此方法也称为平方和分层解构法。在模型中，每一项只针对它前面的那项进行调整。类型 I 平方和常用于:

- 平衡 ANOVA 模型，其中任何主效应在任何一阶交互效应之前指定，任何一阶交互效应在任何双向交互效应之前指定，依此类推。
- 多项式回归模型，其中任何低阶项在任何高阶项之前指定。
- 纯嵌套模型，其中第一个指定的效应嵌套在第二个指定的效应中，第二个指定的效应嵌套在第三个指定的效应中，依此类推。（此嵌套形式只能通过使用语法来指定。）

类型 II。此方法在为所有其它“相应的”效应进行调节的模型中计算某个效应的平方和。相应的效应是指，与所有效应（不包含正被检查的效应）相对应的效应。类型 II 平方和法常用于:

- 平衡 ANOVA 模型。
- 任何只有主要因子效应的模型。
- 任何回归模型。
- 纯嵌套设计。（此嵌套形式能通过使用语法来指定。）

类型 III。缺省类型。此方法按以下方式计算设计中某项效应的平方和: 作为针对任何其他不包含该效应的效应进行调整，并与任何包含该效应的效应（如果存在）正交的平方和。类型 III 平方和具有一个主要优点，那就是只要可估计性的一般形式保持不变，平方和对于单元格频率就保持不变。因此，我们常认为此类平方和对于不带缺失单元格的不平衡模型有用。在不带缺失单元格的因子设计中，此方法等同于 Yates 加权均方方法。类型 III 平方和法常用于:

- 任何在类型 I 和类型 II 中列出的模型。
- 任何不带空白单元格的平衡或非平衡模型。

类型 IV。此方法针对存在缺失单元格的情况设计。对于设计中的任何效应 *F*，如果任何其它效应中不包含 *F*，那么类型 IV = 类型 III = 类型 II。当 *F* 包含在其它效应中时，那么类型 IV 将 *F* 中的参数中正在进行的对比相等地分配到所有较高水平的效应。类型 IV 平方和法常用于:

- 任何在类型 I 和类型 II 中列出的模型。
- 任何带有空白单元格的平衡或非平衡模型。

GLM 对比

对比用来检验因子的水平之间的差值。您可以为模型中的每个因子指定对比（在重复测量模型中，那么是为每个主体间因子）。对比代表参数的线性组合。

GLM 单变量。假设检验基于原假设 **LB** = 0，其中 **L** 是对比系数矩阵，**B** 是参数向量。在指定对比时，将创建 **L** 矩阵。对应于因子的 **L** 矩阵列与对比匹配。对剩余的列进行调整，使 **L** 矩阵可以估计。

输出包含每组对比的 *F* 统计。为对比差值显示的还有基于 Student *t* 分布的 Bonferroni 型同时置信区间。

可用对比

可用对比有偏移对比、简单对比、差分对比、Helmert 对比、重复对比和多项式对比。对于偏移对比和简单对比，您可以选择参考类别是最后一个类别还是第一个类别。

对比类型

偏差。将每个水平（参考类别除外）的平均值与所有水平的平均值（总平均值）进行比较。因子的水平可以为任何顺序。

简单。将每个水平的平均值与指定水平的平均值进行比较。当存在控制组时，此类对比很有用。可以选择第一个或最后一个类别作为参考类别。

差分。将每个水平的平均值（第一个水平除外）与前面水平的平均值进行比较。（有时候称为逆 Helmert 对比。）

Helmert。将因子的每个水平的平均值（最后一个水平除外）与后面水平的平均值进行比较。

重复。将每个水平的平均值（最后一个水平除外）与后一个水平的平均值进行比较。

多项式。比较线性效应、二次效应、三次效应等等。第一自由度包含跨所有类别的线性效应；第二自由度包含二次效应，依此类推。这些对比常常用来估计多项式趋势。

GLM 概要图

概要图（交互图）对于比较模型中的边际平均值是有用的。概要图是一个线图，其中每个点表示因子的一个水平上的估计因变量边际平均值（已针对任何协变量进行调整）。第二个因子的水平可用来绘制分离线。第三个因子中的每个水平可用来创建分离图。所有固定和随机因子（如果存在）可用于图。对于多变量分析，将为每个因变量创建概要图。在重复测量分析中，主体间因子和主体内因子均可以用在概要图中。只有在安装了“Advanced Statistics”选项的情况下，“GLM 多变量”和“GLM 重复测量”才可用。

单因子的概要图显示估计边际平均值是沿水平增加还是减小。对于两个或更多因子，平行线表示因子之间没有交互，这意味着您只能调查一个因子的水平。不平行的线则表示交互。

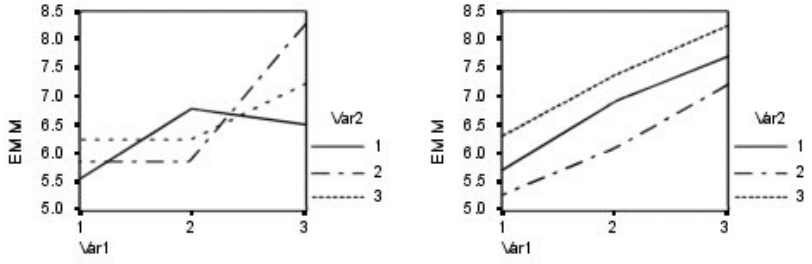

图 2. 不平行图（左）和平行图（右）

在通过为水平轴选择因子，以及通过为分离线和分离图选择因子（后者可选）指定了图之后，该图必须添加到“图”列表中。

GLM: 选项

此对话框中有一些可选统计。统计是使用固定效应模型计算的。

估计边际平均值。选择您需要的单元格中的总体边际平均值估计的因子和交互作用。为协变量（如果存在）调整这些平均值。

- **比较主效应**。对于主体间和主体内因子，为模型中的任何主效应提供估计边际平均值未修正的成对比较。只有在“显示以下项的平均值”列表中选择了主效应的情况下，此项才可用。

- **置信区间调节**。选择最小显著性差异 (LSD)、Bonferroni 或对置信区间和显著性的 Sidak 调整。此项只有在选择了**比较主作用**的情况下才可用。

输出。选择**描述统计**以生成所有单元格中的所有因变量的观察到的平均值、标准差和计数。**功效估计**给出了每个作用和每个参数估计值的偏 eta 方值。eta 方统计描述总可变性中可归因于某个因子的部分。当基于观察值设置备用假设时，选择**观察势**可获取检验的势。选择**参数估计**可为每个检验生成参数估计值、标准误差、*t* 检验、置信区间和检验的观察势。选择**对比系数矩阵**可获取 **L** 矩阵。

同质性检验为跨主体间因子所有水平组合的每个因变量生成 Levene 的方差同质性检验（仅对于主体间因子）。分布-水平图和残差图选项对于检查关于数据的假设很有用。如果不存在任何因子，那么禁用此项。选择**残差图**可为每个因变量生成观察-预测-标准化残差图。这些图对于调查方差相等的假设很有用。选择**失拟**可检查因变量和自变量之间的关系是否能由模型充分地描述。**常规可估计函数**允许您基于常规可估计函数构造定制的假设检验。任何对比系数矩阵中的行均是常规可估计函数的线性组合。

显著性水平。您可能想要调整用在事后检验中的显著性水平，以及用于构造置信区间的置信度。指定的值还用于计算检验的观察势。如果指定了显著性水平，那么相关联的置信区间度会显示在对话框中。

UNIANOVA 命令的附加功能

使用命令语法语言还可以:

- 在设计中指定嵌套效应（使用 DESIGN 子命令）。
- 指定效应对比效应的线性组合或一个值的检验（使用 TEST 子命令）。
- 指定多个对比（使用 CONTRAST 子命令）。
- 包括用户缺失值（使用 MISSING 子命令）。
- 指定 EPS 标准（使用 CRITERIA 子命令）。
- 构造定制的 **L** 矩阵、**M** 矩阵或 **K** 矩阵（使用 LMATRIX、MMATRIX 和 KMATRIX 子命令）。
- 为偏移对比或简单对比指定中间参考类别（使用 CONTRAST 子命令）。
- 为多项式对比指定矩阵（使用 CONTRAST 子命令）。
- 为事后比较指定误差项（使用 POSTHOC 子命令）。
- 为因子列表中的任何因子或因子之间的因子交互计算估计边际平均值（使用 EMMEANS 子命令）。
- 为临时变量指定名称（使用 SAVE 子命令）。
- 构造相关性矩阵数据文件（使用 OUTFILE 子命令）。
- 构造包含主体间 ANOVA 表中的统计的矩阵数据文件（使用 OUTFILE 子命令）。
- 将设计矩阵保存到新的数据文件（使用 OUTFILE 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

GLM 事后比较

事后多重比较检验。一旦确定平均值间存在差值，两两范围检验和成对多重比较就可以确定哪些平均值存在差值了。对未调整的值进行比较。这些检验只用于固定的主体间因子。在“GLM 重复测量”中，如果没有主体间因子，那么这些检验不可用，但为跨主体内因子的水平的平均值执行事后多重比较检验。对于“GLM 多变量”，那么分别为每个因变量执行事后检验。只有在安装了“Advanced Statistics”选项的情况下，“GLM 多变量”和“GLM 重复测量”才可用。

Bonferroni 和 Tukey's 真实显著性差异检验是常用的多重比较检验。**Bonferroni 检验**基于 Student 的 *t* 统计，它针对已进行多重比较这一事实调整观察的显著性水平。**Sidak 的 t 检验**也调整显著性水平，并提供比

Bonferroni 检验更严密的界限。**Tukey's 真实显著性差异检验**使用 Student 化的范围统计量在组之间进行所有成对比较，并将试验误差率设置为所有成对比较的集合的误差率。当检验大量平均值对时，Tukey's 真实显著性差异检验比 Bonferroni 检验更有效。对于少量的对，Bonferroni 更有效。

Hochberg's GT2 类似于 Tukey 真实显著性差异检验，但使用了 Student 化的最大值模数。通常 Tukey 的检验更有效。**Gabriel 的成对比较检验**也使用 Student 化的最大值模数，在单元格尺寸不等的情况下通常比 Hochberg's GT2 更有效。当单元格大小变化过大时，Gabriel 检验可能会变得随意。

Dunnett 的成对多重比较 t 检验将一组处理与单个控制平均值进行比较。最后一个类别是缺省的控制类别。另外，您还可以选择第一个类别。您还可以选择双侧或单尾检验。要检验因子的任何水平（控制类别除外）的平均值是否不等于控制类别的平均值，请使用双侧检验。要检验因子的任何水平的平均值是否小于控制类别的平均值，请选择 **< 控制**。类似地，要检验因子的任何水平的平均值是否大于控制类别的平均值，请选择 **> 控制**。

Ryan、Einot、Gabriel 和 Welsch (R-E-G-W) 开发了两个多重逐步降低范围检验。多重逐步降低过程首先检验所有平均值是否相等。如果不是所有的平均值均相等，那么检验一部分平均值的等同性。**R-E-G-W F** 基于 *F* 检验，而 **R-E-G-W Q** 基于 Student 化的范围。这些检验要比 Duncan 的多范围检验和 Student-Newman-Keuls（也是多重逐步下降过程）有效，但对于不相等的单元格大小则不推荐使用它们。

当方差不等时，使用 **Tamhane's T2**（基于 *t* 检验的保守成对比较检验）、**Dunnett T3**（基于 Student 化的最大模数的成对比较检验）、**Games-Howell 成对比较检验**（有时是随意的）或者 **Dunnett's C**（基于 Student 化的范围的成对比较检验）。注意，如果模型中有多个因子，这些检验无效且不会被生成。

Duncan 的多范围检验、Student-Newman-Keuls (**S-N-K**) 和 **Tukey 的 b** 是排列组平均值等级的范围检验，并计算范围值。这些检验的使用频率不如先前讨论的检验。

Waller-Duncan t 检验使用 Bayesian 方法。当样本大小不相等时，此范围检验使用样本大小的调和平均值。

Scheffé 检验的显著性水平设计成允许检验组平均值的所有可能线性组合，而不仅仅允许此功能中可用的成对比较。其结果是，Scheffé 检验常常比其他检验更保守，这意味着显著性要求平均值之间存在更大的差别。

最小显著性差异 (**LSD**) 成对多重比较检验等同于所有组对之间的多重个别 *t* 检验。此检验的缺点是，不进行任何尝试来为多重比较调整观察到的显著性水平。

显示的检验。为 LSD、Sidak、Bonferroni、Games-Howell、Tamhane's T2 和 T3、Dunnett's *C* 以及 Dunnett T3 提供成对比较。为 S-N-K、Tukey 的 *b*、Duncan、R-E-G-W *F*、R-E-G-W *Q* 以及 Waller 提供范围检验的均一子集。Tukey 真实显著性差异检验、Hochberg GT2、Gabriel 检验以及 Scheffé 检验既是多重比较检验，同时也是范围检验。

GLM: 选项

此对话框中有一些可选统计。统计是使用固定效应模型计算的。

估计边际平均值。选择您需要的单元格中的总体边际平均值估计的因子和交互作用。为协变量（如果存在）调整这些平均值。

- **比较主效应**。对于主体间和主体内因子，为模型中的任何主效应提供估计边际平均值未修正的成对比较。只有在"显示以下项的平均值"列表中选择了主效应的情况下，此项才可用。
- **置信区间调节**。选择最小显著性差异 (LSD)、Bonferroni 或对置信区间和显著性的 Sidak 调整。此项只有在选择了**比较主作用**的情况下才可用。

输出。选择**描述统计**以生成所有单元格中的所有因变量的观察到的平均值、标准差和计数。**功效估计**给出了每个作用和每个参数估计值的偏 eta 方值。eta 方统计描述总可变性中可归因于某个因子的部分。当基于观察值设

置备用假设时，选择**观察势**可获取检验的势。选择**参数估计**可为每个检验生成参数估计值、标准误差、*t* 检验、置信区间和检验的观察势。选择**对比系数矩阵**可获取 **L** 矩阵。

同质性检验为跨主体间因子所有水平组合的每个因变量生成 Levene 的方差同质性检验（仅对于主体间因子）。分布-水平图和残差图选项对于检查关于数据的假设很有用。如果不存在任何因子，那么禁用此项。选择**残差图**可为每个因变量生成观察-预测-标准化残差图。这些图对于调查方差相等的假设很有用。选择**失拟**可检查因变量和自变量之间的关系是否能由模型充分地描述。**常规可估计函数**允许您基于常规可估计函数构造定制的假设检验。任何对比系数矩阵中的行均是常规可估计函数的线性组合。

显著性水平。您可能想要调整用在事后检验中的显著性水平，以及用于构造置信区间的置信度。指定的值还用于计算检验的观察势。如果指定了显著性水平，那么相关联的置信区间度会显示在对话框中。

UNIANOVA 命令的附加功能

使用命令语法语言还可以:

- 在设计中指定嵌套效应（使用 DESIGN 子命令）。
- 指定效应对比效应的线性组合或一个值的检验（使用 TEST 子命令）。
- 指定多个对比（使用 CONTRAST 子命令）。
- 包括用户缺失值（使用 MISSING 子命令）。
- 指定 EPS 标准（使用 CRITERIA 子命令）。
- 构造定制的 **L** 矩阵、**M** 矩阵或 **K** 矩阵（使用 LMATRIX、MMATRIX 和 KMATRIX 子命令）。
- 为偏移对比或简单对比指定中间参考类别（使用 CONTRAST 子命令）。
- 为多项式对比指定矩阵（使用 CONTRAST 子命令）。
- 为事后比较指定误差项（使用 POSTHOC 子命令）。
- 为因子列表中的任何因子或因子之间的因子交互计算估计边际平均值（使用 EMMEANS 子命令）。
- 为临时变量指定名称（使用 SAVE 子命令）。
- 构造相关性矩阵数据文件（使用 OUTFILE 子命令）。
- 构造包含主体间 ANOVA 表中的统计的矩阵数据文件（使用 OUTFILE 子命令）。
- 将设计矩阵保存到新的数据文件（使用 OUTFILE 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

GLM: 保存

您可以在数据编辑器中将模型预测的值、残差和相关测量另存为新变量。这些变量中有许多可用于检查关于数据的假设。要保存供另一 IBM SPSS Statistics 会话中使用的值，您必须保存当前数据文件。

预测值。模型为每个个案预测的值。

- *未标准化*。模型为因变量预测的值。
- *加权*。加权未标准化预测值。仅在之前已选择了 WLS 变量的情况下可用。
- *标准误差*。对于自变量具有相同值的个案所对应的因变量的平均值的标准差的估计。

诊断。标识以下个案的测量：自变量的值具有不寻常组合的个案，以及可能对模型产生很大影响的个案。

- *Cook 距离*。在特定个案从回归系数的计算中排除的情况下，所有个案的残差变化幅度的测量。较大的 Cook 距离表明从回归统计的计算中排除个案之后，系数会发生根本变化。
- *杠杆值*。未居中的杠杆值。每个观察值对模型拟合度的相对影响。

残差。未标准化残差是因变量的实际值减去由模型预测的值。还提供标准化残差、Student 化的残差以及剔除残差。如果选择了 WLS 变量，那么提供加权的未标准化残差。

- *未标准化*。观察值与模型预测值之间的差。
- *加权*。加权未标准化残差。仅在之前已选择了 WLS 变量的情况下可用。
- *标准化*。残差除以其标准差的估计。标准化残差也称为 Pearson 残差，它的平均值为 0，标准差为 1。
- *Student 化*。残差除以其随个案变化的标准差的估计，这取决于每个个案的自变量值与自变量平均值之间的距离。
- *剔除*。当某个案从回归系数的计算中排除时，该个案的残差。它是因变量的值和调整预测值之间的差。

系数统计。将模型中的参数估计值的协方差矩阵写入当前会话中的新数据集，或写入外部 IBM SPSS Statistics 数据文件。而且，对于每个因变量，将存在一行参数估计值、一行与参数估计值对应的 *t* 统计的显著性值以及一行残差自由度。对于多变量模型，每一个因变量都存在类似的行。您可以在读取矩阵文件的其他过程中使用此矩阵文件。

GLM: 选项

此对话框中有一些可选统计。统计是使用固定效应模型计算的。

估计边际平均值。选择您需要的单元格中的总体边际平均值估计的因子和交互作用。为协变量（如果存在）调整这些平均值。

- **比较主效应**。对于主体间和主体内因子，为模型中的任何主效应提供估计边际平均值未修正的成对比较。只有在“显示以下项的平均值”列表中选择了主效应的情况下，此项才可用。
- **置信区间调节**。选择最小显著性差异 (LSD)、Bonferroni 或对置信区间和显著性的 Sidak 调整。此项只有在选择了**比较主作用**的情况下才可用。

输出。选择**描述统计**以生成所有单元格中的所有因变量的观察到的平均值、标准差和计数。**功效估计**给出了每个作用和每个参数估计值的偏 eta 方值。eta 方统计描述总可变性中可归因于某个因子的部分。当基于观察值设置备用假设时，选择**观察势**可获取检验的势。选择**参数估计**可为每个检验生成参数估计值、标准误差、*t* 检验、置信区间和检验的观察势。选择**对比系数矩阵**可获取 **L** 矩阵。

同质性检验为跨主体间因子所有水平组合的每个因变量生成 Levene 的方差同质性检验（仅对于主体间因子）。分布-水平图和残差图选项对于检查关于数据的假设很有用。如果不存在任何因子，那么禁用此项。选择**残差图**可为每个因变量生成观察-预测-标准化残差图。这些图对于调查方差相等的假设很有用。选择**失拟**可检查因变量和自变量之间的关系是否能由模型充分地描述。**常规可估计函数**允许您基于常规可估计函数构造定制的假设检验。任何对比系数矩阵中的行均是常规可估计函数的线性组合。

显著性水平。您可能想要调整用在事后检验中的显著性水平，以及用于构造置信区间的置信度。指定的值还用于计算检验的观察势。如果指定了显著性水平，那么相关联的置信区间度会显示在对话框中。

UNIANOVA 命令的附加功能

使用命令语法语言还可以:

- 在设计中指定嵌套效应（使用 DESIGN 子命令）。
- 指定效应对比效应的线性组合或一个值的检验（使用 TEST 子命令）。
- 指定多个对比（使用 CONTRAST 子命令）。
- 包括用户缺失值（使用 MISSING 子命令）。
- 指定 EPS 标准（使用 CRITERIA 子命令）。

- 构造定制的 **L** 矩阵、**M** 矩阵或 **K** 矩阵（使用 LMATRIX、MMATRIX 和 KMATRIX 子命令）。
- 为偏移对比或简单对比指定中间参考类别（使用 CONTRAST 子命令）。
- 为多项式对比指定矩阵（使用 CONTRAST 子命令）。
- 为事后比较指定误差项（使用 POSTHOC 子命令）。
- 为因子列表中的任何因子或因子之间的因子交互计算估计边际平均值（使用 EMMEANS 子命令）。
- 为临时变量指定名称（使用 SAVE 子命令）。
- 构造相关性矩阵数据文件（使用 OUTFILE 子命令）。
- 构造包含主体间 ANOVA 表中的统计的矩阵数据文件（使用 OUTFILE 子命令）。
- 将设计矩阵保存到新的数据文件（使用 OUTFILE 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 12 章 双变量相关性

双变量相关性过程计算 Pearson 相关性系数、Spearman 的 rho 和 Kendall 的 tau-*b* 及其显著性水平。相关性测量变量或等级顺序的相关方式。在计算相关系数之前，先过滤数据以找出离群值（离群值可能会导致误导性的结果）和线性关系的证据。Pearson 相关性系数是一种线性相关性测量。两个变量可能良好相关，但是如果其关系不是线性的，那么 Pearson 相关性系数就不是适合度量其相关性的统计。

示例。一个篮球队所赢得的比赛次数与其每场比赛所得的平均分数相关吗？散点图表明，两者间存在线性关系。对 1994-1995 NBA 赛季数据的分析结果是，Pearson 相关性系数 (0.581) 的显著性水平为 0.01。您可能会猜想，每个赛季所赢得的比赛越多，对手所得的分数就越少。这些变量是负相关的 (-0.401)，并且相关显著性水平为 0.05。

统计。对于每个变量：具有非缺失值的个案数、平均值以及标准差。对于每对变量：Pearson 相关性系数、Spearman Rho、Kendall tau-*b*、偏差的叉积以及协方差。

双变量相关性数据注意事项

数据。对 Pearson 相关性系数使用对称的定量变量，对 Spearman 的 rho 和 Kendall 的 tau-*b* 使用定量变量或具有已排序类别的变量。

假设。Pearson 相关性系数假定每对变量是二元正态分布。

获取双变量相关性

从菜单中选择:

分析 > 关联 > 双变量...

1. 选择两个或更多数值型变量。

还可以使用以下选项:

- **相关系数。**对于正态分布的定量变量，请选择 **Pearson** 相关系数。如果您的数据不是正态分布的，或具有已排序的类别，请选择 **Kendall 的 tau-b** 或 **Spearman**，后两者度量等级顺序之间的相关性。相关系数的值范围为 -1（完全负相关）到 +1（完全正相关）。0 值表示没有线性关系。在解释结果时请小心谨慎，不要因显著的相关性而得出任何因果结论。
- **显著性检验。**您可以选择双尾概率或单尾概率。如果预先已知关联的方向，请选择**单尾**。否则，请选择**双尾**。
- **标记显著性相关**。用一个星号来标识显著性水平为 0.05 的相关系数，用两个星号来标识显著性水平为 0.01 的相关系数。

双变量相关性选项

统计。对于 Pearson 相关性，您可以选择以下一项或两项:

- **平均值和标准差**。为每个变量显示。还显示具有非缺失值的个案数。无论缺失值设置如何，都将逐变量处理缺失值。
- **叉积偏差和协方差**。为每对变量显示。偏差的叉积等于校正平均值变量的乘积之和。这是 Pearson 相关性系数的分子。协方差是有关两个变量之间的关系的一种非标准化度量，等于叉积偏差除以 *N*-1。

© Copyright IBM Corporation 1989, 2014

缺失值。您可以选择以下选项之一:

- **按对排除个案**。会从分析中排除对其计算相关系数的一对变量中一个或两个含有缺失值的个案。由于每个系数均基于对特定变量对具有有效代码的所有个案，因此在每次计算中会使用可用的最大信息量。这可能因为个案数不同而产生一组系数。
- **按列表排除个案**。从所有相关性中排除对任意变量有缺失值的个案。

CORRELATIONS 和 NONPAR CORR 命令的附加功能

使用命令语法语言还可以:

- 写入 Pearson 相关性的相关性矩阵，用其替代原始数据，从而获取诸如因子分析之类的其他分析（使用 MATRIX 子命令）。
- 获取一个列表上的每个变量与另一个列表上的每个变量的相关性（在 VARIABLES 子命令中使用关键字 WITH）。

请参阅*命令语法参考*以获取完整的语法信息。

第 13 章 偏相关

“偏相关”过程计算偏相关系数，该系数在控制一个或多个附加变量的效应的同时描述两个变量之间的线性关系。相关是线性相关性测量。两个变量可以完全相关，但如果关系不是线性的，那么相关系数就不是适合度量它们相关性的统计。

示例。在保健基金和发病率之间存在关系吗？尽管您可能希望此类关系都是负相关关系，但研究表明存在显著的正相关关系：随着保健基金的增长，发病率也表现为增长。不过，对保健提供商的拜访率的控制，实际上消除了所观察到的正相关。保健基金和发病率显示为正相关的原因仅仅是：当基金增长时，更多的人可以获得保健服务，从而导致医生和医院所报告的病例更多。

统计。对于每个变量：具有非缺失值的个案数、平均值以及标准差。偏相关性矩阵和零阶相关性矩阵，以及自由度和显著性水平。

偏相关数据注意事项

数据。使用对称的定量变量。

假设。“偏相关”过程假定每对变量都是二元正态的。

获取偏相关

1. 从菜单中选择:

   分析 > 关联 > 偏...
2. 选择两个或更多要为之计算偏相关的数值变量。
3. 选择一个或多个数值控制变量。

还可以使用以下选项:

- **显著性检验。**您可以选择双尾概率或单尾概率。如果预先已知关联的方向，请选择**单尾**。否则，请选择**双尾**。
- **显示实际显著性水平**。缺省情况下，将显示每个相关系数的概率和自由度。如果取消选择此项，那么使用单个星号标识显著性水平为 0.05 的系数，使用两个星号标识显著性水平为 0.01 的系数，而不显示自由度。此设置同时影响偏相关性矩阵和零阶相关性矩阵。

偏相关: 选项

统计。可以选择以下方式中的一个或两个都选:

- **平均值和标准差**。为每个变量显示。还显示具有非缺失值的个案数。
- **零阶相关系数**。显示所有变量（包括控制变量）之间简单相关的矩阵。

缺失值。您可以选择以下选项之一:

- **按列表排除个案**。将从所有计算中排除其任何变量（包括控制变量）具有缺失值的个案。
- **按对排除个案**。对于偏相关所基于的零阶相关的计算，不使用其一对变量或其中一个变量具有缺失值的个案。按对删除可以充分使用数据。但是，个案数可能随系数的不同而不同。如果按对删除有效，那么某个特定的偏相关系数的自由度是基于在任何零阶相关计算中使用的最小个案数。

© Copyright IBM Corporation 1989, 2014

PARTIAL CORR 命令的附加功能

使用命令语法语言还可以:

- 读取零阶相关性矩阵或写入偏相关性矩阵（使用 MATRIX 子命令）。
- 获取两个变量列表之间的偏相关（使用 VARIABLES 子命令上的关键字 WITH）。
- 获取多个分析（使用多个 VARIABLES 子命令）。
- 有两个控制变量时，指定请求的阶数值（例如，第一阶和第二阶偏相关）（使用 VARIABLES 子命令）。
- 排除冗余系数（使用 FORMAT 子命令）。
- 当无法计算某些系数时，显示简单相关的矩阵（使用 STATISTICS 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 14 章 距离

此过程计算测量变量对或个案对之间相似性或非相似性（距离）的各种统计。随后，这些相似性或距离测量可与其他过程（例如因子分析、聚类分析或多维刻度）一起使用，以帮助分析复杂的数据集。

示例。有可能基于某些特征（例如引擎大小、MPG 和马力）度量汽车对之间的相似性？通过计算汽车间的相似性，您可以了解到哪些汽车彼此相似，哪些汽车彼此不同。对更正式的分析，您可以考虑将分层聚类分析或多维刻度应用到相似性中，以探索基础结构。

统计。定距数据的非相似性（距离）度量有欧氏距离、平方 Euclidean 距离、Chebychev、块、Minkowski 或定制度量；计数数据的非相似性（距离）度量有卡方或 phi 平方；二分类数据的非相似性（距离）度量有欧氏距离、平方 Euclidean 距离、刻度差分、模式差分、方差、形状或 Lance 和 Williams。定距数据的相似性测量有 Pearson 相关性或余弦；二分类数据的相似性测量有 Russel 和 Rao、简单匹配、Jaccard、切块、Rogers 和 Tanimoto、Sokal 和 Sneath 1、Sokal 和 Sneath 2、Sokal 和 Sneath 3、Kulczynski 1、Kulczynski 2、Sokal 和 Sneath 4、Hamann、Lambda、Anderberg 的 *D*、Yule 的 *Y*、Yule 的 *Q*、Ochiai、Sokal 和 Sneath 5、phi 4 点相关或离差。

获得距离矩阵

1. 从菜单中选择:

   分析 > 关联 > 距离...

2. 选择至少一个数值变量来计算个案间的距离，或选择至少两个数值变量来计算变量间的距离。
3. 在“计算距离”组中选择一个选项来计算个案或变量间的近似性。

距离: 非相似性测量

从“度量”组中选择与数据类型（区间、计数或二值）相应的选项，然后，在下拉列表中选择与该数据类型相应的测量。根据数据类型，可用的测量有:

- **定距数据**。欧氏距离、平方 Euclidean 距离、Chebychev、块、Minkowski 或定制。
- **计数数据**。卡方测量或 phi 平方测量。
- **二分类数据**。欧氏距离、平方 Euclidean 距离、刻度差分、模式差分、方差、形状或 Lance 和 Williams。（在“存在”和“不存在”中输入值以指定哪两个值有意义，“距离”将忽略其他所有值。）

“转换值”组允许您在计算近似性之*前*，为个案或变量标准化数据值。对二分类数据，这些转换不适用。可用的标准化方法包括 *z* 得分、范围从 -1 到 1、范围从 0 到 1、最大量级为 1、平均值为 1 以及标准差为 1。

“转换测量”组允许您转换距离测量所生成的值。在计算了距离测量之后应用这些转换。可用的选项包括绝对值、更改符号以及重定比例到 0-1 范围。

距离: 相似性测量

从“度量”组中选择与数据类型（定距或二分类）相应的选项，然后，在下拉列表中选择与该数据类型相应的测量。根据数据类型，可用的测量有:

- **定距数据**。Pearson 相关性或余弦。

© Copyright IBM Corporation 1989, 2014

- **二分类数据**。Russell 和 Rao、简单匹配、Jaccard、切块、Rogers 和 Tanimoto、Sokal 和 Sneath 1、Sokal 和 Sneath 2、Sokal 和 Sneath 3、Kulczynski 1、Kulczynski 2、Sokal 和 Sneath 4、Hamann、Lambda、Anderberg 的 *D*、Yule 的 *Y*、Yule 的 *Q*、Ochiai、Sokal 和 Sneath 5、phi 4 点相关或离差。（在“存在”和“不存在”中输入值以指定哪两个值有意义，“距离”将忽略其他所有值。）

“转换值”组允许您在计算近似性之前，为个案或变量标准化数据值。对二分类数据，这些转换不适用。可用的标准化方法包括 *z* 得分、范围从 -1 到 1、范围从 0 到 1、最大量级为 1、平均值为 1 以及标准差为 1。

“转换测量”组允许您转换距离测量所生成的值。在计算了距离测量之后应用这些转换。可用的选项包括绝对值、更改符号以及重定比例到 0-1 范围。

PROXIMITIES 命令的附加功能

“距离”过程使用 PROXIMITIES 命令语法。使用命令语法语言还可以:

- 指定任意整数作为 Minkowski 距离测量的幂。
- 指定任意整数作为定制距离测量的幂和根。

请参阅*命令语法参考*以获取完整的语法信息。

第 15 章 线性模型

线性模型根据目标与一个或多个预测变量间的线性关系来预测连续目标。

线性模型相对简单，用于评分的数学公式也易于解释。这些模型的属性比较好理解，与同一数据集上的其他模型类型（如神经网络或决策树）相比能够非常快速构建。

示例。在调查业主保险理赔方面拥有有限资源的保险公司希望构建一个模型来估计理赔成本。通过在服务中心部署该模型，客服代表可以在接听客户电话的同时输入理赔信息，并立即获得基于以往数据的“预期”成本。请参阅主题以获取更多信息。

字段要求。必须有一个目标和至少一个输入。缺省情况下，不使用带“两者”或“无”预定义角色的字段。目标必须为连续（刻度）。对预测变量（输入）没有测量级别限制。分类（名义、有序）字段用作模型中的因子，同时连续字段用作协变量。

注意: 如果分类字段包含 1000 个以上的类别，那么不会运行此过程，也不会构建任何模型。

要获取线性模型

此功能需要 Statistics Base 选项。

从菜单中选择:

分析 > 回归 > 自动线性模型...

1. 确保至少有一个目标和一个输入。
2. 单击**构建选项**以指定可选的构建与模型设置。
3. 单击**模型选项**以保存得分到活动数据集并导出模型到外部文件。
4. 单击**运行**以运行过程并创建模型对象。

目标

您的主要目标是什么？ 请选择适当的目标。

- **创建一个标准模型**。使用一种方法来构建一个可以使用预测变量预测目标的模型。一般来说，标准模型更易于理解，而且评分速度比 boosted、bagged 或大型数据集整体更快。
- **增强模型精确性 (Boosting)**。使用 Boosting 构建整体模型的方法，可生成一个模型序列来获得更多精确预测值。与标准模型相比，整体模型需要更长的时间来构建和评分。

  Boosting 将生成一连串的“组件模型”，其中的每个模型都基于整个数据集进行构建。在构建每个连续的组件模型之前，将根据上一个组件模型的残值确定记录的权重。残值较大的个案将被赋予相对较高的分析权重，以使下一个组件模型还侧重于预测这些记录。这些组件模型共同构成整体模型。这个整体模型使用组合规则对新记录进行评分；可用的规则取决于目标的测量级别。
- **增强模型稳定性 (Bagging)**。使用 Bagging (bootstrap 汇总) 构建整体模型的方法，可生成多个模型来获得更多可靠的预测值。与标准模型相比，整体模型需要更长的时间来构建和评分。

© Copyright IBM Corporation 1989, 2014

Bootstrap 汇总 (Bagging) 通过对原始数据集进行放回方式的取样，生成训练数据集的副本。这将创建大小与原始数据集相同的 Bootstrap 样本。然后，以每个副本为基础构建“组件模型”。这些组件模型共同构成整体模型。这个整体模型使用组合规则对新记录进行评分；可用的规则取决于目标的测量级别。

- **为超大型数据集创建模型（需要 IBM SPSS Statistics Server）。**通过将数据集拆分成单独的数据块来构建整体模型的方法。如果您的数据集非常大，无法构建任何上述模型，或希望用于增量建模，请选择该选项。该选项可使用更短的时间来构建模型，但需要比标准模型更长的时间来评分。此选项需要 IBM SPSS Statistics Server 连接。

要了解 Boosting、Bagging 及特大型数据集的相关设置，请参阅第 55 页的『整体』。

基本

自动准备数据。该选项允许在内部转换目标和预测变量，以使模型的预测能力最大化；将保存模型的任何转换并应用到新数据用于评分。转换字段的原始版本将从模型中排除。缺省情况下，执行以下自动数据准备。

- **日期与时间处理。**每个日期预测变量被转换成新的连续预测变量，其中包含自参考日期 (1970-01-01) 以来经过的时间。每个时间预测变量被转换成新的连续预测变量，其中包含自参考时间 (00:00:00) 以来经过的时间。
- **调整测量级别。**具有少于 5 个不同值的连续预测变量将被重新强制转换为有序预测变量。具有多于 10 个不同值的有序预测变量将被重新强制转换为连续预测变量。
- **离群值处理。**如果连续预测变量的值位于分界值（平均值的 3 个标准差）之外，那么将其设为分界值。
- **缺失值处理**。名义预测变量的缺失值被替换为训练分区的众数。有序预测变量的缺失值被替换为训练分区的中位数。连续预测变量的缺失值被替换为训练分区的平均值。
- **受监督的合并。**这将减少与目标关联的需处理的字段数，得到更简约的模型。通过输入与目标间的关系可以确定类似的类别。无显著差异（即 p 值大于 0.1）的类别则被合并。如果所有类别合并为一个类别，那么字段的原始和派生版本将从模型中排除，因为它们没有作为预测变量的值。

置信度。此为用于在系数视图中计算模型系数的区间估计值的置信度。指定大于 0 且小于 100 的值。缺省值为 95。

模型选择

模型选择方法。选择一种模型选择方法（下面将详细介绍）或**包括所有预测变量**，后者简单地输入所有可用预测变量作为主效应模型项。缺省使用**前向逐步**。

前向逐步选择。在开始时模型中没有任何效应，然后在每个步骤中添加和删除效应，直到根据逐步选择标准不能再添加或删除效应为止。

- **纳入/移除标准**。此为用于决定是将某个效应添加到还是剔除出模型的统计。**信息标准 (AICC)** 基于模型中给定训练集合的似然估计，并可调整以惩罚过度复杂模型。**F 统计**基于有关模型错误改进情况的某个统计检验。**调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。**防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

  如果选择了 **F 统计**以外的标准，那么在每步中将对应于选择标准的最大正增长的效应添加到模型。对应于标准中减少情况的任何模型效应将被移除。

  如果选择了 **F 统计**作为标准，那么在每步中将具有低于指定阈值（**纳入 p 值小于此值的效应**）的最小 *p* 值的效应添加到模型。缺省值为 0.05。任何具有大于指定阈值**移除 p 值大于此值的效应**的 *p* 值的模型效应将被移除。缺省值为 .10。

- **自定义最终模型中的最大效应数**。缺省情况下，所有可用效应都将被输入模型中。或者，如果逐步选择算法在具有指定最大效应数的某个步骤结束，那么此算法将以当前效应集合结束。
- **自定义最大步骤数**。逐步选择算法在达到特定步骤数后停止。此值缺省为可用效应数的 3 倍。或者，指定一个正整数作为最大步骤数。

最佳子集选择。这将检查“所有可能的”模型，或至少检查可能模型的较大子集（大于“前向逐步”方法），以选择满足相应标准的最佳子集。 **信息标准 (AICC)** 基于模型中给定训练集合的似然估计，并可调整以惩罚过度复杂模型。**调整 R 方**基于训练集合的拟合度，并可调整以惩罚过度复杂模型。**防止过度拟合准则 (ASE)** 基于防止过度拟合集的拟合度（平均方差，或 ASE）。防止过度拟合集是不用于训练模型且大约为原始数据集 30% 的随机子样本。

选择具有最大标准值的模型作为最佳模型。

注意: 与向前逐步选择相比，最佳子集选择涉及更密集的计算。在与 Boosting、Bagging 或超大型数据集配合执行最佳子集时，花费的时间比使用向前逐步选择构建标准模型要长得多。

整体

这些设置决定了在“目标”中请求 Boosting、Bagging 或超大型数据集时发生的整体行为。对选定目标不适用的选项将被忽略。

Bagging 和大型数据集 在对整体评分时，此规则用于组合来自基本模型的预测值，以计算整体得分值。

- **连续目标的缺省组合规则**。可以通过对来自基本模型的预测值取平均值或中位数，对连续目标的整体预测值进行组合。

注意，如果以增强模型精确性为目标，那么组合规则选择将被忽略。Boosting 方法始终使用加权大多数投票来对分类目标进行评分，而使用加权中位数对连续目标进行评分。

Boosting 和 Bagging。当以增强模型精确性或稳定性为目标时，指定要构建的基本模型数；对于 Bagging 方法，此为 bootstrap 样本数。它应为正整数。

高级

重复结果。设置随机种子允许您复制分析。随机数生成器用于选择哪个记录在过度拟合集中。指定一个整数，或单击**生成**，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。缺省值为 54752075。

模型选项

保存预测值到数据集。缺省变量名称是 *PredictedValue*。

导出模型。这将写入模型到外部 *.zip* 文件。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。指定有效的唯一文件名。如果文件规范引用了现有文件，那么该文件将被覆盖。

模型摘要

“模型摘要”视图是模型及其拟合的快照概览摘要。

表。该表标识一些高级模型设置，包括:

- 字段选项卡上指定的目标名称,
- 是否已按基本设置中指定的方式执行执行自动数据准备,
- 模型选择设置中指定的模型选择方法和选择标准。还显示了最终模型的选择标准值，并以较小、较佳的格式显示。

图表。此图表显示最终模型的精确性，数值越大越好。对于最终模型，此值为 100 × 调整后的 R^2。

自动数据准备

此视图显示在自动数据准备 (ADP) 步骤中排除了哪些字段，以及转换字段的派生方式等信息。对于每个转换或排除字段，在此表中列出了字段名、在分析中的角色，以及 ADP 步骤所采取的操作。这些字段按其名称的字母升序排列。对每个字段可能执行的操作包括:

- **派生持续时间: 月份**计算从包含日期的字段中的值到当前系统日期经过的时间（以月份为单位）。
- **派生持续时间: 小时**计算从包含时间的字段中的值到当前系统时间经过的时间（以小时为单位）。
- **将测量级别从连续改为有序**将不到 5 个唯一值的连续字段重新强制转换为有序字段。
- **将测量级别从有序改为连续**将超过 10 个唯一值的有序字段重新强制转换为连续字段。
- **删除离群值**如果连续预测变量的值位于分界值（平均值的 3 个标准差）之外，那么将其设为分界值。
- **替换缺失值**分别使用众数、中位数和平均值替换名义字段、有序字段和连续字段的缺失值。
- **合并类别以最大化与目标的关联**根据输入与目标间的关系确定“类似”的预测变量类别。无显著差异（即 p 值大于 0.05）的类别则被合并。
- **排除常量预测变量/在离群值处理之后/在合并类别之后**删除具有单个值的预测变量，可能在执行其他 ADP 操作之后。

预测变量重要性

通常，您需要将建模工作专注于最重要的预测变量字段，并考虑删除或忽略那些最不重要的预测变量字段。预测变量重要性图表可以在模型估计中指示每个预测变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有预测变量的值总和为 1.0。预测变量重要性与模型精度无关。它只与每个预测变量在预测中的重要性有关，而不涉及预测是否精确。

按已观测进行预测

这将显示一个分级散点图，其中预测值位于垂直轴上，而观测值位于水平轴上。理想情况下，该点应在 45 度线上；您可以从该视图上判断出任何被模型预测为较差的纪录。

残差

这将显示模型残差的诊断图表。

图表样式。有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **直方图。**此为 Student 化的残差的分级直方图，并带有正态分布交叠。线性模型假设残差具有正态分布，因此理想情况下直方图应相当接近平滑线。
- **P-P 图。**此为分级概率-概率 (P-P) 图，将 Student 化的残差与正态分布进行对比。如果绘制点的斜率比正态线更平缓，那么残差显示出比正态分布更显著的可变性；如果更陡峭，那么残差的可变性低于正态分布。如果绘制点呈 S 型曲线，那么残差为偏斜分布。

离群值

此表列出对模型施加过度影响的记录，并显示记录 ID（如果在“字段”选项卡上指定）、目标值，以及 Cook 距离。Cook 距离是在特定记录从模型系数的计算中排除的情况下，所有记录的残差变化幅度的测量。较大的 Cook 距离表示在排除记录后系数会发生显著变化，因此应被视为有一定影响。

应仔细检查有影响的记录，以确定是在模型估计中给予较低权重，按照特定可接受阈值截断离群值，还是彻底移除有影响的记录。

效应

此视图显示模型中每个效应的大小。

样式。有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **图表。**在此图表中，将按预测变量重要性递减顺序，从上到下排列显示效应。在图表中，连接线条根据效应的显著性进行加权，粗线条表示较显著的效应（*p* 值较小）。悬停在连接线条上将显示工具提示，以指示效应的 *p* 值和重要性。这是缺省值。
- **表**。此为总体模型与单独模型效应的 ANOVA 表。各个效应将按预测变量重要性递减顺序，从上到下排列显示。注意，在缺省情况下，此表处于折叠状态，只显示总体模型结果。要查看单独模型效应的结果，在表中单击**校正的模型**单元格。

预测变量重要性。提供有一个“预测变量重要性”滑块，以控制在视图中显示哪些预测变量。这不会改变模型，只是帮助您重点关注最重要的预测变量。缺省显示前 10 个效应。

显著性。提供有一个“显著性”滑块，以便在按预测变量重要性显示效应的基础上，进一步控制在视图中显示哪些效应。显著性值大于滑块值的效应将被隐藏。这不会改变模型，只是帮助您重点关注最重要的效应。缺省情况下此值为 1.00，因此不会根据显著性来过滤效应。

系数

此视图显示模型中每个系数的值。注意，由于因子（分类预测变量）在模型内部经过指示符编码，因此包含因子的**效应**通常具有多个关联**系数**；每种类别一个关联系数，但对应于冗余（参考）参数的类别除外。

样式。有多种不同的显示样式，可以从**样式**下拉列表中访问这些样式。

- **图表。**在此图表中，首先显示截距，然后按预测变量重要性递减顺序，从上到下排列显示效应。在包含因子的效应中，系数按照数据值的升序进行排列。在图表中，连接线条根据系数的显著性（参见图表键）而具有不同颜色，粗线条表示较显著的系数（*p* 值较小）。悬停在连接线条上将显示工具提示，以指示与参数关联的效应的系数值、*p* 值和重要性。这是缺省样式。
- **表**。这将显示单独模型系数的值、显著性检验，以及置信区间。在截距后面，各个效应将按预测变量重要性递减顺序，从上到下排列显示。在包含因子的效应中，系数按照数据值的升序进行排列。注意，在缺省情况下，此表处于折叠状态，只显示每个模型参数的系数、显著性和重要性。要查看标准误差、*t* 统计和置信区

间，在表中单击**系数**单元格。悬停在表中的模型参数名称上，将显示工具提示，以指示参数名称、与参数关联的效应以及与模型参数关联的值标签（对于分类预测变量）。当自动数据准备合并分类预测变量的相似类别时，这尤其适合用于查看新创建的类别。

预测变量重要性。提供有一个“预测变量重要性”滑块，以控制在视图中显示哪些预测变量。这不会改变模型，只是帮助您重点关注最重要的预测变量。缺省显示前 10 个效应。

显著性。提供有一个“显著性”滑块，以便在按预测变量重要性显示系数的基础上，进一步控制在视图中显示哪些系数。显著性值大于滑块值的系数将被隐藏。这不会改变模型，只是帮助您重点关注最重要的系数。缺省情况下此值为 1.00，因此不会根据显著性来过滤系数。

估计平均值

只为显著的预测变量显示这些图表。在图表中，目标的模型估计值位于垂直轴上，预测变量的每个值位于水平轴上，所有其他预测变量保持恒定。它提供了有关每个预测变量系数在目标上的效应的可视化，非常有用。

注意: 如果没有显著的预测变量，那么不会生成估计平均值。

模型构建摘要

如果在“模型选择”设置中选择了**无**以外的模型选择算法，这将提供有关模型构建过程的一些详细信息。

前向逐步。如果选择算法为前向逐步，此表将显示逐步选择算法中的最近 10 步。对于其中每个步骤，显示在此步骤上选择标准的值与模型中的效应。这允许您了解每个步骤对模型的贡献大小。每列允许您对行进行排序，因此可以方便地看到在给定步骤上模型中有哪些效应。

最佳子集。如果选择算法为最佳子集，此表将显示前 10 个模型。对于每个模型，显示选择标准的值与模型中的效应。您可以从中了解这些最佳模型的稳定性；如果它们倾向于具有存在少量差异的相似效应，那么您可以充分确信它们的确是“最佳”模型；如果它们倾向于具有迥异的效应，那么某些效应可能太相似，需要进行合并（或删除一些）。每列允许您对行进行排序，因此可以方便地看到在给定步骤上模型中有哪些效应。

第 16 章 线性回归

“线性回归”估计包含一个或多个自变量的线性方程的系数，这些系数能最佳地预测因变量的值。例如，可尝试根据诸如年龄、教育程度和从业年数等自变量来预测销售人员的全年销售额（因变量）。

示例。某个篮球队在一个赛季中获胜的场次与该队在每场比赛中的平均得分相关吗？散点图表示这些变量线性相关。获胜的场次与对手的平均得分也是线性相关的。这些变量负相关。随着获胜场次的增加，对手的平均得分减少。通过线性回归，您可以对这些变量的关系进行建模。好的模型可以用来预测球队的获胜场次。

统计。对于每个变量：有效个案数、平均值和标准差。对于每个模型：回归系数、相关性矩阵、部分相关和偏相关、复 R、R^2、调整 R^2、R^2 变化、估计值的标准误差、方差分析表、预测值和残差。另外还包括：每个回归系数的 95% 置信区间、方差-协方差矩阵、方差膨胀因子、容差、Durbin-Watson 检验、距离测量（Mahalanobis、Cook 和杠杆值）、DfBeta、DfFit、预测区间和个案诊断信息。图：散点图、部分图、直方图和正态概率图。

线性回归数据注意事项

数据。因变量和自变量必须是定量的。分类变量（例如宗教、主要研究领域或居住地）需要记录到二分类（哑元）变量或其他类型的对比变量中。

假设。对于自变量的每个值，因变量必须呈正态分布。对于自变量的所有值，因变量分布的方差必须是恒定的。因变量和每个自变量之间的关系应是线性的，且所有观察值应是独立的。

获取线性回归分析

1. 从菜单中选择:

   分析 > 回归 > 线性...

2. 在“线性回归”对话框中，选择一个数值型因变量。
3. 选择一个或多个数值型自变量。

根据需要，您可以:

- 将自变量分组成块，并对不同的变量子集指定不同的进入方法。
- 选择一个选择变量，将分析限于包含此变量特定值的个案子集。
- 选择个案标识变量，用于标识图上的点。
- 选择数值型 WLS 权重变量以进行加权最小二次方分析。

WLS. 允许您获取加权最小二次方模型。以数据点方差的倒数对数据点进行加权。这意味着方差较大的观察值对分析的影响比方差较小的观察值要小。如果加权变量的值为 0、负数或缺失，那么将该个案从分析中排除。

线性回归变量选择方法

方法选择允许您指定自变量将如何进入到分析中。通过使用不同的方法，您可以从相同的变量组构造多个回归模型。

- *输入（回归）(Enter (Regression)).* 一种变量选择过程，其中一个块中的所有变量在一个步骤中输入。

© Copyright IBM Corporation 1989, 2014

- *逐步式 (Stepwise).* 在每一步，不在方程中的具有 F 的概率最小的自变量被选入（如果该概率足够小）。对于已在回归方程中的变量，如果它们的 F 概率变得足够大，那么移去这些变量。如果不再有变量符合包含或移去的条件，那么该方法终止。
- *移除 (Remove).* 一种变量选择过程，其中在单步中移去一个块中的所有变量。
- *向后去除 (Backward Elimination).* 一种变量选择过程，在该过程中将所有变量输入到方程中，然后按顺序移去。会考虑将与因变量之间的部分相关性最小的变量第一个移去。如果它满足消除条件，那么将其移去。移去第一个变量之后，会考虑下一个将方程的剩余变量中具有最小的部分相关性的变量移去。直到方程中没有满足消除条件的变量，过程才结束。
- *向前选择 (Forward Selection).* 一个逐步变量选择过程，在该过程中将变量顺序输入到模型中。第一个考虑要选入到方程中的变量是与因变量之间具有最大的正或负的相关性的变量。只要在该变量满足选入条件时才将它选入到方程中。选入了第一个变量之后，接下来考虑不在方程中的具有最大的部分相关性的自变量。当无满足选入条件的变量时，过程结束。

输出中的显著性值基于与单个模型的拟合。所以，当使用步进法（逐步式、向前或向后）时，显著性值通常无效。

无论指定什么进入方法，所有变量都必须符合容差条件才能进入方程。缺省的容差水平为 0.0001。另外，如果某个变量会导致另一已在模型中的变量的容差下降到容差条件以下，那么该变量不进入方程。

所有被选自变量将被添加到单个回归模型中。不过，您可以为不同的变量子集指定不同的进入方法。例如，您可以使用逐步式选择将一个变量块输入到回归模型中，而使用向前选择输入第二个变量块。要将第二个变量块添加到回归模型，请单击**下一个**。

线性回归: 设置规则

分析中包含由选择规则定义的个案。例如，如果选择变量，选择**等于**，并为该值键入 5，那么只有那些选定变量值等于 5 的个案才会包含在分析中。字符串值也是允许的。

线性回归: 图

图可以帮助验证正态性、线性相关度和方差相等的假设。对于检测离群值、异常观察值和有影响的个案，图也是有用的。将它们保存为新变量之后，在数据编辑器中可以使用预测值、残差和其他诊断信息来构造含有自变量的图。以下图是可用的:

散点图。您可以对下列各项中的任意两项进行绘图: 因变量、标准化预测值、标准化残差、剔除残差、调整预测值、Student 化残差或者 Student 化剔除残差。针对标准化预测值绘制标准化残差，以检查线性相关度和等方差性。

源变量列表。列出因变量 (DEPENDNT) 及以下预测变量和残差变量: 标准化预测值 (*ZPRED)、标准化残差 (*ZRESID)、剔除残差 (*DRESID)、调整的预测值 (*ADJPRED)、Student 化的残差 (*SRESID) 以及 Student 化的已删除残差 (*SDRESID)。

生成所有部分图。当根据其余自变量分别对两个变量进行回归时，显示每个自变量残差和因变量残差的散点图。要生成部分图，方程中必须至少有两个自变量。

标准化残差图。您可以获取标准化残差的直方图和正态概率图，将标准化残差的分布与正态分布进行比较。

如果请求了任意图，那么将显示标准化预测值和标准化残差（***ZPRED** 和 ***ZRESID**）的汇总统计。

线性回归: 保存新变量

您可以保存预测值、残差和其他对于诊断信息有用的统计。每选择一次将向活动数据文件添加一个或多个新变量。

预测值。回归模型对每个个案预测的值。

- *未标准化*。模型为因变量预测的值。
- *标准化*。每个预测值转换为其标准化形式的转换。即，预测值减去平均值预测值，得到的差除以预测值的标准差。标准化预测值的平均值为 0，标准差为 1。
- *调节*。当某个案从回归系数的计算中排除时，个案的预测值。
- *平均值预测值的标准误差*。预测值的标准误差。对于自变量具有相同值的个案所对应的因变量的平均值的标准差的估计。

距离。标识以下个案的测量：自变量的值具有异常组合的个案，以及可能对回归模型产生很大影响的个案。

- *Mahalanobis*。自变量上个案的值与所有个案的平均值相异程度的测量。大的马氏距离表示个案在一个或多个自变量上具有极值。
- *Cook*。在特定个案从回归系数的计算中排除的情况下，所有个案的残差变化幅度的测量。较大的 Cook 距离表明从回归统计的计算中排除个案之后，系数会发生根本变化。
- *杠杆值*。度量某个点对回归拟合的影响。集中的杠杆值范围为从 0（对拟合无影响）到 (N-1)/N。

预测区间。平均值和个别预测区间的上界和下界。

- *平均值 (Mean)*. 平均预测响应的预测区间的下限和上限（两个变量）。
- *单值*。单个个案的因变量预测区间的下限和上限（两个变量）。
- *置信区间*。输入 1 到 99.99 之间的值，以指定两个预测区间的置信度。在输入此值之前必须选择“平均值”或“区间”。典型的置信区间值为 90、95 和 99。

残差。因变量的实际值减去按回归方程预测的值。

- *未标准化*。观察值与模型预测值之间的差。
- *标准化*。残差除以其标准差的估计。标准化残差也称为 Pearson 残差，它的平均值为 0，标准差为 1。
- *Student 化*。残差除以其随个案变化的标准差的估计，这取决于每个个案的自变量值与自变量平均值之间的距离。
- *剔除*。当某个案从回归系数的计算中排除时，该个案的残差。它是因变量的值和调整预测值之间的差。
- *Student 化剔除*。个案的剔除残差除以其标准误差。Student 化的剔除残差与其相关联的 Student 化的残差之间的差分指示去除某个个案对其预测产生的差分。

影响统计。由于排除了特定个案而导致的回归系数 (DfBeta) 和预测值 (DfFit) 的变化。标准化 DfBeta 和 DfFit 值也可与协方差比率一起使用。

- *DfBeta(s)* . beta 值的差分是由于排除了某个特定个案而导致的回归系数的改变。为模型中的每一项（包括常数项）均计算一个值。
- *标准化 DfBeta (Standardized DfBeta)*. Beta 值的标准化差分。由于排除了某个特定个案而导致的回归系数的改变。您可能想要检查除以 N 的平方根之后绝对值大于 2 的个案，其中 N 是个案数。为模型中的每一项（包括常数项）均计算一个值。
- *DfFit*. 拟合值的差分是由于排除了某个特定个案而产生的预测变量的更改。
- *标准化 DfFit (Standardized DfFit)*. 拟合值的标准化差分。由于排除了某个特定个案而导致的预测值的改变。您可能想要检查绝对值大于 p/N 的平方根的 2 倍的标准化值，其中 p 是模型中的参数个数，N 是个案数。

- *协方差比率*。从回归系数计算中排除特定个案的协方差矩阵的行列式与包含所有个案的协方差矩阵的行列式的比率。如果比率接近 1，那么说明被排除的个案不能显著改变协方差矩阵。

系数统计。将回归系数保存到数据集或数据文件。可以在同一会话中继续使用数据集，但不会将其另存为文件，除非在会话结束之前明确将其保存为文件。数据集名称必须符合变量命名规则。

将模型信息输出到 XML 文件。将参数估计值及其（可选）协方差导出到指定的 XML (PMML) 格式的文件。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。

线性回归: 统计

可用统计有:

回归系数。 **估计**显示回归系数 *B*、*B*的标准误差、标准化系数Beta、*B* 的 *t* 值以及 *t* 的双尾显著性水平。**置信区间**显示具有每个回归系数或协方差矩阵的指定置信度的置信区间。**协方差矩阵**显示回归系数的方差-协方差矩阵，其对角线以外为协方差，对角线上为方差。还显示相关性矩阵。

模型拟合度。列出输入到模型中的变量以及从模型中除去的变量，并显示以下拟合优度统计: 复 *R*、R^2 和调整 R^2、估计的标准误差以及方差分析表。

R 方变化。由于添加或删除自变量而产生的 R^2 统计的更改。如果与某个变量相关联的 R^2 变化很大，那么意味着该变量是因变量的一个良好的预测变量。

描述性。提供分析中的有效个案数、平均值以及每个变量的标准差。还显示具有单尾显著性水平的相关性矩阵以及每个相关系数的个案数。

偏相关 (Partial Correlation). 对于两个变量，在移去由于它们与其他变量之间的相互关联引起的相关性之后，这两个变量之间剩余的相关性。对于因变量与某个自变量，当已移去模型中的其他自变量对上述两者的线性效应之后，这两者之间的相关性。

部分相关 (Part Correlation). 对于因变量与某个自变量，当已移去模型中的其他自变量对该自变量的线性效应之后，因变量与该自变量之间的相关性。当变量添加到方程时，它与 R 方的更改有关。有时称为半部分相关。

共线性诊断。由于一个自变量是其他自变量的线性函数时所引起的共线性（或多重共线性）是不被期望的。显示已标度和未中心化交叉积矩阵的特征值、条件指数以及方差-分解比例，以及个别变量的方差膨胀因子 (VIF) 和容差。

残差。显示残差的序列相关性的 Durbin-Watson 检验，以及满足选择标准（*n* 倍标准差以外的离群值）的个案的个案诊断信息。

线性回归: 选项

可用选项有:

步进法标准。这些选项在已指定向前、向后或逐步式变量选择法的情况下适用。变量可以进入到模型中，或者从模型中移去，这取决于 *F* 值的显著性（概率）或者 *F* 值本身。

- *使用 F 的概率 (Use Probability of F).* 如果变量的 F 值的显著性水平小于“输入”值，那么将该变量选入到模型中，如果该显著性水平大于“剔除”值，那么将该变量从模型中移去。“输入”值必须小于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请增加“输入”值。要将更多的变量从模型中移去，请降低“剔除”值。

- *使用 F 的值 (Use F Value).* 如果变量的 F 值大于“输入”值，那么该变量输入模型，如果 F 值小于“剔除”值，那么该变量从模型中移去。“输入”值必须大于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请降低“输入”值。要将更多的变量从模型中移去，请增大“剔除”值。

在等式中包含常量。缺省情况下，回归模型包含常数项。取消选择此选项可强制使回归通过原点，实际上很少这样做。某些通过原点的回归结果无法与包含常数的回归结果相比较。例如，不能以通常的方式解释 R^2。

缺失值。您可以选择以下选项之一:

- **按列表排除个案**。只有所有变量均取有效值的个案才包含在分析中。
- **按对排除个案**。使用正被相关的变量对具有完整数据的个案来计算回归分析所基于的相关系数。自由度基于最小成对 *N*。
- **使用平均值替换**。将所有个案用于计算，用变量的平均值替换缺失观察值。

REGRESSION 命令的附加功能

使用命令语法语言还可以:

- 写入相关性矩阵或读取矩阵代替原始数据，以获取回归分析（使用 MATRIX 子命令）。
- 指定容差水平（使用 CRITERIA 子命令）。
- 获取相同或不同因变量的多个模型（使用 METHOD 和 DEPENDENT 子命令）。
- 获取其他统计（使用 DESCRIPTIVES 和 STATISTICS 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 17 章 序数回归

使用序数回归可以在一组预测变量（可以是因子或协变量）上对多歧分序数响应的依赖性进行建模。序数回归的设计基于 McCullagh (1980, 1998) 的方法论；序数回归的过程在语法中称为 PLUM。

标准线性回归分析是指最小化响应变量（因变量）和预测变量（自变量）的加权组合之间的平方和差值。估计的系数反映了预测变量的变化对响应的影响程度。假设响应是数值，在此意义上，在整个响应范围内响应水平的变化是等同的。例如，身高为 150 厘米的人和身高为 140 厘米的人的高度差为 10 厘米，该高度差的含义与身高为 210 厘米的人和身高为 200 厘米的人的高度差的含义相同。这些关系不一定适用于序数变量，序数变量中响应类别的选择和数量可能会非常随意。

示例。可以使用序数回归研究患者对药物剂量的反应。可能的反应可以分为 *无*、*轻微*、*适度*或*剧烈*。轻微反应和适度反应之间的差别很难或不可能量化，并且这种差别是取决于感觉的。另外，轻微反应和适度反应之间的差别可能比适度反应和剧烈反应之间的差别更大或更小。

统计和图。观察的和期望的频率以及累积频率、频率和累积频率的 Pearson 残差、观察到的和期望的概率、观察到的和期望的以协变量模式表示的每个响应类别的累积概率、参数估计值的渐近相关性和协方差矩阵、Pearson 的卡方和似然比卡方统计、拟合优度统计、迭代历史记录、平行线假定的检验、参数估计值、标准误差、置信区间以及 Cox 和 Snell、Nagelkerke 和 McFadden 的 R^2 统计。

序数回归数据注意事项

数据。假设因变量是序数并且可以是数值或字符串。通过对因变量的值进行升序排序来确定排列顺序。最低值定义第一个类别。假设因子变量是分类变量。协变量必须为数值。请注意：使用多个连续协变量很容易使创建的单元格概率表非常大。

假设。只允许使用一个响应变量，并且必须指定该响应变量。另外，对于多个自变量值的各个不同模式，假设该响应是独立的多项变量。

相关过程。标定 Logistic 回归对于名义因变量使用相似的模型。

获取序数回归

1. 从菜单中选择:

   分析 > 回归 > 有序...
2. 选择一个因变量。
3. 单击**确定**。

序数回归: 选项

使用“选项”对话框可以调整迭代估计算法中所使用的参数，选择参数估计值的置信度并选择关联函数。

迭代。可以定制迭代算法。

- **最大迭代次数**。指定一个非负整数。如果指定为 0，那么过程会返回初始估计值。
- **最大步骤对分**。指定一个正整数。
- **对数似然估计收敛性性**。如果对数似然估计中的绝对或相对变化小于该值，那么算法会停止。如果指定 0，那么不使用该条件。

© Copyright IBM Corporation 1989, 2014

- **参数收敛**。如果每个参数估计值中的绝对或相对变化小于该值，那么算法会停止。如果指定 0，那么不使用该条件。

置信区间。指定一个大于等于 0 且小于 100 的值。

Delta。添加到零单元格频率的值。指定一个小于 1 的非负值。

奇异性容差。用于检查具有高度依赖性的预测变量。从选项列表中选择一个值。

关联函数。关联函数是累积概率的转换形式，可用于模型估计。下列 5 个关联函数可用。

- **Logit**。$f(x)=\log(x/(1-x))$。通常用于均匀分布的类别。
- **互补双对数**。$f(x)=\log(-\log(1-x))$。通常在较高类别的可能性更大的情况下使用。
- **负双对数**。$f(x)=-\log(-\log(x))$。通常在较低类别的可能性更大的情况下使用。
- **概率**。$f(x)=\Phi^{-1}(x)$。通常在潜在变量正态分布的情况下使用。
- **Cauchit（逆 Cauchy）**。$f(x)=\tan(\pi(x-0.5))$。通常在潜在变量具有许多极值的情况下使用。

序数回归输出

使用“输出”对话框可以生成在查看器中显示的表，并将变量保存到工作文件。

输出。为以下项目生成表:

- **打印迭代历史记录**。为所指定的打印迭代频率打印对数似然估计和参数估计值。始终打印第一个和最后一个迭代。
- **拟合优度统计**。Pearson 和似然比卡方统计。基于在变量列表中指定的分类计算这些统计。
- **汇总统计**。Cox 和 Snell、Nagelkerke 和 McFadden R^2 统计。
- **参数估计**。参数估计值、标准误差和置信区间。
- **参数估计的渐近相关性**。参数估计相关系数的矩阵。
- **参数估计的渐近协方差**。参数估计协方差的矩阵。
- **单元格信息**。观察的和期望的频率和累积频率、频率和累积频率的 Pearson 残差、观察到的和期望的概率以及以协变量模式表示的观察到的和期望的每个响应类别的累积概率。请注意：对于具有许多协变量模式的模型（例如，具有连续协变量的模型），该选项可能会生成非常大的、很难处理的表。
- **平行线检验**。位置参数在多个因变量水平上都相等的假设检验。该检验只对仅定位模型可用。

保存的变量。将以下变量保存到工作文件:

- **估计响应概率**。将因子/协变量模式分类成响应类别的模型估计概率。概率与响应类别的数量相等。
- **预测类别**。具有因子/协变量模式的最大估计概率的响应类别。
- **预测类别概率**。将因子/协变量分类成预测类别的估计概率。该概率也是因子/协变量模式的估计概率的最大值。
- **实际类别概率**。将因子/协变量分类成实际类别的估计概率。

打印对数似然性。控制对数似然估计的显示。**包含多项式常数**可以提供似然估计的完整值。若要在不包含该常数的乘积之间比较结果，可以选择将该常数排除。

序数回归: 位置模型

使用“位置”对话框可以指定分析的位置模型。

指定模型。主效应模型包含协变量和因子的主效应，但不包含交互效应。可以创建自定义模型以指定因子交互效应或协变量交互效应的子集。

因子/协变量。列出因子与协变量。

位置模型。该模型取决于所选择的主效应和交互效应。

建立项

对于选定因子和协变量:

交互。创建所有选定变量的最高级交互项。这是缺省值。

主效应。为每个选定的变量创建主效应项。

所有二阶。创建选定变量的所有可能的双向交互。

所有三阶。创建选定变量的所有可能的三阶交互。

所有四阶。创建选定变量的所有可能的四阶交互。

所有五阶。创建选定变量的所有可能的五阶交互。

序数回归: 刻度模型

使用“度量”对话框可以指定分析的刻度模型。

因子/协变量。列出因子与协变量。

度量模型。该模型取决于所选择的主效应和交互效应。

建立项

对于选定因子和协变量:

交互。创建所有选定变量的最高级交互项。这是缺省值。

主效应。为每个选定的变量创建主效应项。

所有二阶。创建选定变量的所有可能的双向交互。

所有三阶。创建选定变量的所有可能的三阶交互。

所有四阶。创建选定变量的所有可能的四阶交互。

所有五阶。创建选定变量的所有可能的五阶交互。

PLUM 命令的附加功能

如果将您的选择粘贴到语法窗口并编辑结果 PLUM 命令语法，那么可以定制序数回归。使用命令语法语言还可以:

- 通过将原假设指定为参数的线性组合来创建定制的假设检验。

请参阅*命令语法参考*以获取完整的语法信息。

第 18 章 曲线估计

曲线估计过程为 11 种不同的曲线估计回归模型生成曲线估计回归统计和相关的图。将对每个因变量生成一个单独的模型。也可以将预测值、残差和预测区间保存为新变量。

示例。一个 Internet 服务提供商跟踪其网络上随时间变化的受病毒感染的电子邮件流量百分比。散点图显示关系是非线性的。您可以用二次或立方模型来拟合数据，并检查假设的有效性和模型的拟合优度。

统计。对于每个模型：回归系数、复 R、R^2、调整 R^2、估计值的标准误差、方差分析表、预测值、残差和预测区间。模型：线性、对数、逆、二次、三次、幂、复合、S 曲线、Logistic、增长和指数。

曲线估计数据注意事项

数据。因变量和自变量必须是定量的。如果从活动数据集中选择**时间**作为自变量（而不是选择变量），那么曲线估计过程生成个案之间时间长度均匀的时间变量。如果选择了**时间**，那么该因变量应为时间序列度量。时间序列分析需要这样一种数据文件结构：其中每个个案（行）代表不同时间的一组观察值，而个案之间的时间长度是均匀的。

假设。以图形方式过滤数据，以确定自变量和因变量的相关方式（线性相关、指数相关等）。好模型的残差应呈随机正态分布。如果使用了线性模型，那么应满足下列假设：对于自变量的每个值，因变量必须呈正态分布。对于自变量的所有值，因变量分布的方差必须是恒定的。因变量和自变量之间的关系应该为线性关系，而所有观察值应该是独立的。

获取曲线估计

1. 从菜单中选择:

   分析 > 回归 > 曲线估计...
2. 选择一个或多个因变量。将对每个因变量生成一个单独的模型。
3. 选择一个自变量（选择活动数据集中的变量或选择**时间**）。
4. 或者:

- 选择一个变量以用于在散点图中标注个案。对于散点图中的每个点，您可以使用“点选择”工具来显示个案标签变量的值。
- 单击**保存**将预测值、残差和预测区间保存为新变量。

还可以使用以下选项:

- **在等式中包含常量**。估计回归方程式中的常数项。缺省情况下包含常数。
- **根据模型绘图**。对照自变量绘制因变量的值和每个选定的模型。为每个因变量产生一个单独的图表。
- **显示 ANOVA 表格**。为每个选定的模型显示摘要方差分析表。

曲线估计：模型

您可以选择一个或多个曲线估计回归模型。要确定使用哪种模型，请绘制数据。如果变量显示为线性相关，那么使用简单线性回归模型。当变量不是线性相关时，请尝试转换数据。当转换没有帮助时，那么可能需要更复杂的模型。查看数据的散点图；如果该图看起来像是您了解的某个数学函数，那么将数据与该类型的模型进行拟合。例如，如果数据看起来像指数函数，请使用指数模型。

© Copyright IBM Corporation 1989, 2014

线性。方程为 Y = b0 + (b1 * t) 的模型。按时间的线性函数建模的序列值。

对数。方程为 Y = b0 + (b1 * ln(t)) 的模型。

逆。方程为 Y = b0 + (b1 / t) 的模型。

二次。方程式为 Y = b0 + (b1 * t) + (b2 * t**2) 的模型。二次模型可用来对“减弱”的序列或阻尼衰减的序列进行建模。

三次。由方程 Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3) 定义的模型。

幂。方程式为 Y = b0 * (t**b1) 或 ln(Y) = ln(b0) + (b1 * ln(t)) 的模型。

复合。方程为 Y = b0 * (b1**t) 或 ln(Y) = ln(b0) + (ln(b1) * t) 的模型。

S 曲线。方程式为 Y = e**(b0 + (b1/t)) 或 ln(Y) = b0 + (b1/t) 的模型。

Logistic。方程式为 Y = 1 / (1/u + (b0 * (b1**t))) 或 ln(1/y-1/u) = ln (b0) + (ln(b1) * t) 的模型，其中 u 是上边界值。选择“逻辑”之后，请指定用在回归方程中使用的上界值。该值必须是一个大于最大因变量值的正数。

增长。方程式为 Y = e**(b0 + (b1 * t)) 或 ln(Y) = b0 + (b1 * t) 的模型。

指数。方程为 Y = b0 * (e**(b1 * t)) or ln(Y) = ln(b0) + (b1 * t) 的模型。

曲线估计: 保存

保存变量。对于每个选定的模型，您可以保存预测值、残差（因变量的观察值减去模型预测值）和预测区间（上限和下限）。新变量名称和描述标签显示在输出窗口中的表中。

预测个案。在活动数据集中，如果选择**时间**而不是变量作为自变量，那么可以指定超出时间序列结尾的预测期。您可以选择以下选项之一:

- **从估计期到最后一个个案的预测**。在估计期内的个案的基础上预测文件中所有个案的值。显示在对话框底端的估计期可通过“数据”菜单上的“选择个案”选项的“范围”子对话框来定义。如果未定义任何估计期，那么使用所有个案来预测值。
- **预测范围**。根据估计期中的个案，预测指定日期、时间或观察号范围内的值。此功能可以用于预测超出时间序列中最后一个个案的值。当前定义的日期变量确定可用于指定预测期结尾的文本框。如果没有已定义的日期变量，那么您可以指定结尾的观察（个案）号。

使用“数据”菜单上的“定义日期”选项来创建日期变量。

第 19 章 部分最小二次方回归

部分最小二次方回归过程估计部分最小二次方（PLS，也称为“投影到潜在结构”）回归模型。PLS 是一种预测方法，可替代普通最小二乘 (OLS) 回归法、典型相关分析或结构化方程建模，当预测变量高度相关或预测变量数量超过个案数目时，此方法尤其有用。

PLS 融合主成分分析和多重回归功能。它首先提取一组充分解释自变量和因变量之间的协方差的潜在因子。然后，回归步骤使用自变量分解来预测因变量的值。

表。解释方差比例（潜在因子）、潜在因子权重、潜在因子加载、图像自变量重要性 (VIP) 和回归参数估计值（因变量）全部缺省生成。

图表。前三个潜在因子的变量投影重要性 (VIP)、因子得分和因子权重，以及与模型的距离均根据选项选项卡生成。

部分最小二次方回归数据注意事项

测量级别。因变量和自（预测）变量可以是刻度、名义或有序变量。此过程假定已对所有变量指定相应的测量级别，尽管您可以右键单击源变量列表中的变量并从弹出菜单中选择测量级别，以临时更改变量的测量级别。此过程以相同的方式处理类别（名义或有序）变量。

类别变量编码。此过程在其间使用一个 *c* 编码临时对类别因变量重新编码。如果存在变量的 *c* 类别，那么变量存储为 *c* 向量，第一个类别指示为 (1,0,...,0)，下一个类别 (0,1,0,...,0)，...，和最后一个类别 (0,0,...,0,1)。使用虚拟编码表示类别因变量，即仅省略对应于参考类别的指示符。

频率权重。权重值在使用前四舍五入为最接近的整数。在分析中不使用缺失权重或权重小于 0.5 的个案。

缺失值。用户和系统缺失值视为无效。

重定比例。所有模型变量均被居中和标准化，包括表示类别变量的指示符变量。

获取部分最小二次方回归

从菜单中选择:

分析 > 回归 > 部分最小二次方...

1. 选择至少一个因变量。
2. 选择至少一个自变量。

根据需要，您可以:

- 指定类别（名义或有序）因变量的参考类别。
- 指定用作个案输出并保存数据集的唯一标识的变量。
- 指定要提取的潜在因子数目的上限。

先决条件

“部分最小二次方回归”过程是一个 Python 扩展命令，并需要缺省情况下随 IBM SPSS Statistics 产品一起安装的 IBM SPSS Statistics - Essentials for Python。另外，它还需要免费提供的 NumPy 和 SciPy Python 库。

© Copyright IBM Corporation 1989, 2014

注: 对于采用分布式分析方式（需要 IBM SPSS Statistics Server）工作的用户，必须在服务器上安装 NumPy 和 SciPy。请与系统管理员联系以获取帮助。

Windows 和 Mac 用户

对于 Windows 和 Mac，必须将 NumPy 和 SciPy 安装到并非随 IBM SPSS Statistics 一起安装的 Python 2.7 版本中。如果您没有另一个版本的 Python 2.7，可以从 http://www.python.org 下载。然后，安装用于 Python V2.7 的 NumPy 和 SciPy。安装程序可以从 http://www.scipy.org/Download 获得。

为了能够使用 NumPy 和 SciPy，必须将 Python 位置设置为安装有 NumPy 和 SciPy 的 Python 2.7 版本。您可以通过“选项”对话框（编辑 > 选项）中的“文件位置”选项卡来设置 Python 位置。

Linux 用户

我们建议您下载源代码并自己构建 NumPy 和 SciPy。源代码可以从 http://www.scipy.org/Download 获得。您可以将 NumPy 和 SciPy 安装到随 IBM SPSS Statistics 一起安装的 Python 2.7 版本中。此版本位于 IBM SPSS Statistics 安装位置下的 `Python` 目录中。

如果您选择将 NumPy 和 SciPy 安装到并非随 IBM SPSS Statistics 一起安装的 Python 2.7 版本中，那么必须将 Python 位置设置为指向该版本。您可以通过“选项”对话框（编辑 > 选项）中的“文件位置”选项卡来设置 Python 位置。

Windows 和 Unix 服务器

在服务器上，必须将 NumPy 和 SciPy 安装到并非随 IBM SPSS Statistics 一起安装的 Python 2.7 版本中。如果服务器上没有另一版本的 Python 2.7，那么可以从 http://www.python.org 进行下载。用于 Python 2.7 的 NumPy 和 SciPy 可以从 http://www.scipy.org/Download 获得。为了能够使用 NumPy 和 SciPy，必须将服务器的 Python 位置设置为安装有 NumPy 和 SciPy 的 Python 2.7 版本。Python 位置可以通过 IBM SPSS Statistics Administration Console 进行设置。

模型

指定模型效应。主效应模型包含所有因子和协变量主效应。选择**定制**指定交互。必须指定要包含在模型中的所有项。

因子与协变量。列出因子与协变量。

模型。模型取决于数据的性质。选择**定制**之后，您可以选择分析中感兴趣的主效应和交互效应。

建立项

对于选定因子和协变量:

交互。创建所有选定变量的最高级交互项。这是缺省值。

主效应。为每个选定的变量创建主效应项。

所有二阶。创建选定变量的所有可能的双向交互。

所有三阶。创建选定变量的所有可能的三阶交互。

所有四阶。创建选定变量的所有可能的四阶交互。

所有五阶。创建选定变量的所有可能的五阶交互。

选项

“选项”选项卡允许用户保存和绘制单个个案、潜在因子和预测变量的模型估计值。

对于每种类型的数据，指定数据集名。数据集名必须是唯一的。如果您指定了现有数据集的名称，那么会替换其内容；否则，将创建新的数据集。

- **保存单个案例估计值**。保存以下个案模型估计值：预测值、残差、潜在因子模型距离和潜在因子分数。它也标示潜在因子分数。
- **保存潜在因子估计值**。保存潜在因子加载和潜在因子权重。它也标示潜在因子权重。
- **保存自变量估计值**。保存回归参数估计值和图像变量重要性（VIP）。它也通过潜在因子标示 VIP。

第 20 章 最近邻元素分析

“最近邻元素分析”方法是根据个案间的相似性来对个案进行分类。在 machine learning 中，它被开发为一种识别数据模式而不需要与任何存储的模式或个案完全匹配的方法。类似个案相互靠近，而不同个案相互远离。因此，通过两个个案之间的距离可以测量他们的非相似性。

相互靠近的个案称为“邻元素”。当出现新个案（保持）时，将计算它与模型中每个个案之间的距离。计算得出最相似个案（最近邻元素）的分类，并将新个案放入包含最多最近邻元素的类别。

您可以指定要检查的最近邻元素数目，该值称为 *k*。

最近邻元素分析也可用于计算连续目标的值。在这种情况下，使用最近邻元素的平均值或中位数目标值来获取新个案的预测值。

最近邻元素分析数据注意事项

目标和特征。目标和特征包括:

- *名义 (Nominal)*. 当变量值表示不具有内在等级的类别时，该变量可以作为名义变量；例如，雇员任职的公司部门。名义变量的示例包括地区、邮政编码和宗教信仰。
- *有序 (Ordinal)*. 当变量值表示带有某种内在等级的类别时，该变量可以作为有序变量；例如，从十分不满意到十分满意的服务满意度水平。有序变量的示例包括表示满意度或可信度的态度分数和优先选择评分。
- *刻度 (Scale)*. 当变量值表示带有有意义的度规的已排序类别时，该变量可以作为刻度（连续）变量对待，以便在值之间进行合适的距离比较。刻度变量的示例包括以年为单位的年龄和以千美元为单位的收入。

  最近邻元素分析以相同的方式处理名义和有序变量。此过程假定已对每个变量指定相应的测量级别；但是，您可以右键单击源变量列表中的变量并从弹出菜单中选择测量级别，以临时更改该变量的测量级别。

变量列表中每个变量旁的图标标识测量级别和数据类型:

表 1. 测量级别图标

	数值	字符串	日期	时间
刻度（连续）		n/a		
有序				
名义				

类别变量编码。此过程使用一个 *c* 编码在过程期间临时重新编码分类预测变量和因变量。如果存在 *c* 分类变量，那么该变量存储为 c 向量，第一个类别表示为（1、0、…、0）、下一个类别表示为（0、1、0、...、0）、...、最后一个类别表示为（0、0、...、0、1）。

此编码方案增加了功能空间的维数。具体来说，维度总数为刻度预测变量数目加上所有分类预测变量间的类别数目。因此，此编码方案可导致训练减速。如果您的最近邻元素训练进行很慢，在运行过程之前，可尝试通过将类似的类别组合起来，或删除具有极少见类别的个案以减少分类预测变量中的类别数目。

所有的“*c* 之一”编码都以训练数据为基础，即使定义了坚持样本也是如此（请参阅第 77 页的『分区』）。因此，如果坚持样本包含训练数据中不存在的预测变量类别个案，那么不对那些个案评分。如果坚持样本包含训练数据中不存在的因变量类别个案，那么对那些个案评分。

重定比例。刻度特征在缺省情况下将标准化。所有重定比例都以训练数据为基础执行，即使定义了坚持样本也是如此（请参阅第 77 页的『分区』）。如果您指定一个变量以定义分区，这些特征在训练样本和坚持样本之间具有相似分布将至关重要。例如，使用探索过程来检查分区间的分布。

频率权重。此过程忽略频率权重。

复制结果。此过程在随机分配分区和交叉验证折期间使用随机数字生成器。如果您希望准确地复制结果，除了使用相同的过程设置以外，还可以为 Mersenne 扭曲器设置种子（请参阅第 77 页的『分区』），或者使用变量来定义分区和交叉验证折。

获取最近邻元素分析

从菜单中选择:

分析 > 分类 > 最近邻元素...

1. 指定一项或多项特征，它们可被视为自变量或预测变量（如果存在目标的话）。

   目标（可选）。如果未指定目标（因变量或响应），那么此过程仅查找 *k* 个最近邻元素 - 而不会执行任何分类或预测。

   标准化刻度特征。标准化特征具有相同的值范围，这可改进估计算法的性能。使用经调整后的标准化 [2*(x-min)/(max-min)]-1。调整后的标准化值介于 -1 与 1 之间。

   焦点个案标识（可选）。这可以标记感兴趣的个案。例如，研究员希望确定一个学区的测验分数（焦点个案）是否与类似学区的测验分数相当。他使用最近邻元素分析来查找在给定特征组方面最相似的学区。然后，他将焦点学区的测验分数与最近邻学区的分数进行比较。

   也可在临床研究中使用焦点个案来选择与临床个案相似的控制个案。焦点个案显示在 *k* 个最近邻元素和距离表、特征空间图表、对等图表和象限图中。有关焦点个案的信息保存到在“输出”选项卡上指定的文件中。

   在指定变量上为正值的个案被视为焦点个案。指定具有非正值的变量是无效的。

个案标签（可选）。在特征空间图表、对等图表和象限图中使用这些值来标记个案。

具有未知测量级别的字段

当数据集中的一个或多个变量（字段）的测量级别未知时，将显示测量级别警告。由于测量级别会影响此过程的计算结果，因此所有变量必须都定义有测量级别。

扫描数据。读取活动数据集中的数据，并分配缺省测量级别给任何具有当前未知测量级别的字段。如果数据集较大，此过程可能需要一些时间。

手动分配。打开列出了所有具有未知测量级别的字段的对话框。您可以使用该对话框将测量级别分配给这些字段。您也可以在数据编辑器的变量视图中分配测量级别。

由于测量级别对此过程很重要，因此您无法访问运行此过程的对话框，除非所有字段均定义了测量级别。

邻元素

最近邻元素的数目 (k)。指定最近邻元素的数目。注意，使用大量的邻元素不一定会得到更准确的模型。

如果在“变量”选项卡中指定了目标，那么可以指定值范围并允许过程选择该范围中“最佳”的邻元素数目。确定最近邻元素数目的方法依赖于“特征”选项卡上要求的特征选择。

- 如果特征选择有效，那么针对请求范围中每个 *k* 值执行特征选择，并选择具有最低误差率（如果目标为刻度，那么为最低平方和误差）的 *k* 值和特征集。
- 如果特征选择未生效，那么使用 *V* 折交叉验证来选择“最佳”的邻元素数目。请参阅“分区”选项卡以控制折指定。

距离计算。该度规用于指定在测量个案相似性中使用的距离度规。

- **Euclidean 度规**。两个个案 x 和 y 之间的距离，为个案值之间的平方差在所有维度上之和的平方根。
- **城市街区度规**。两个个案之间的距离是个案值之间绝对差在所有维度上之和。又称为 Manhattan 距离。

或者，如果在“变量”选项卡上指定了目标，那么可以选择在计算距离时，按标准化的重要性对特征指定权重。预测变量的特征重要性的计算方法为: 不含预测变量的模型的误差率或平方和误差与完整模型的误差率或平方和误差之比。通过重新对特征重要性值指定权重，来计算标准化的重要性，因此其总和为 1。

刻度目标预测。如果在“变量”选项卡上指定了刻度目标，这可指定预测值是基于最近邻元素的平均值还是中值来计算的。

特征

如果在“变量”选项卡中指定了目标，使用“特征”选项卡可以为特征选择请求或指定选项。缺省情况下，特征选择会考虑所有特征，但可以选择特征子集以强制纳入模型。

中止条件。在每一步上，如果添加特征可以使误差最小（计算为分类目标的误差率和刻度目标的平方和误差），那么考虑将其纳入模型中。继续向前选择，直到满足指定的条件。

- **指定的特征数目**。除了那些强制纳入模型的特征外，算法还会添加固定数目的特征。指定一个正整数。减少所选择的数目值可以创建更简约的模型，但存在缺失重要特征的风险。增加所选择的数目值可以涵盖所有重要特征，但又存在因特征添加而增加模型误差的风险。
- **绝对误差比率的最小变化**。当绝对误差比率变化表明无法通过添加更多特征来进一步改进模型时，算法会停止。指定一个正数。减小最小变化值将倾向于包括更多特征，但存在包括对模型价值不大的特征这一风险。增加最小变化值将倾向于排除更多特征，但存在丢失对模型较重要的特征的风险。“最佳”的最小变化值取决于数据和具体应用。请参阅输出中的“特征选择误差日志”，以帮助您评估哪些特征最重要。请参阅 第 81 页的『特征选择误差日志』主题以获取更多信息。

分区

使用“分区”选项卡可以将数据集划分为训练和坚持集，并在适当时候将个案分配给交叉验证折。

训练和坚持分区。此组指定将活动数据集划分为训练样本或坚持样本的方法。**训练样本**包含用于训练最近邻元素模型的数据记录；数据集中的某些个案百分比必须分配给训练样本以获得一个模型。**坚持样本**是用于评估最终模型的独立数据记录集；坚持样本的误差给出一个模型预测能力的“真实”估计值，因为坚持个案不用于构建模型。

- **随机分配个案到分区**。指定分配给训练样本的个案百分比。其余的分配给坚持样本。

- **使用变量分配个案**。指定一个将活动数据集中的每个个案分配到训练或坚持样本中的数值变量。变量为正值的个案被分配到训练样本中，值为 0 或负值的个案被分配到坚持样本中。具有系统缺失值的个案会从分析中排除。分区变量的任何用户缺失值始终视为有效。

交叉验证折。*V* 折交叉验证用于确定“最佳”邻元素数目。因性能原因，它无法与特征选择结合使用。

交叉验证将样本划分为许多子样本，或折。然后，生成最近邻元素模型，并依次排除每个子样本中的数据。第一个模型基于第一个样本折的个案之外的所有个案，第二个模型基于第二个样本折的个案之外的所有个案，依此类推。对于每个模型，估计其错误的方法是将模型应用于生成它时所排除的子样本。“最佳”最近邻元素数为在折中产生最小误差的数量。

- **随机分配个案到折**。指定应当用于交叉验证的折数。此过程将个案随机分配到折，从 1 编号到 *V*（折数）。
- **使用变量分配个案**。指定一个将活动数据集中的每个个案分配到折中的数值变量。变量必须是 1 到 *V* 之间的数字值。如果此范围中的任何值缺失，且位于任何拆分上（如果拆分文件有效），这将导致误差。

为 Mersenne 扭曲器设置种子。设置种子允许您复制分析。使用此控件类似于将“Mersenne 扭曲器”设为活动生成器并在“随机数生成器”对话框中指定固定起始点，两者的重大差别在于在此对话框中设置种子会保留随机数生成器的当前状态并在分析完成后恢复该状态。

保存

保存的变量名称。自动名称生成确保能保存您的所有工作。无需先删除数据编辑器中保存的变量，自定义名称允许您放弃/替换上一次运行的结果。

要保存的变量

- **预测值或类别**。此操作保存刻度目标的预测值或分类目标的预测类别。
- **预测概率**。此操作保存分类目标的预测概率。针对前 *n* 个类别保存单个变量，其中 *n* 在**要为分类目标保存的最大类别数**控制中指定。
- **训练/坚持分区变量**。如果在“分区”选项卡上将个案随机分配到训练和坚持样本中，这将保存个案被分配到的分区（训练或坚持）的值。
- **交叉验证折变量**。如果在“分区”选项卡上将个案随机分配到交叉验证折中，这将保存个案被分配到的折的值。

输出

查看器输出

- **个案处理摘要**。显示个案处理摘要表，其通过训练和坚持样本整体总结分析中包含和排除的个案数。
- **图表和表**。显示模型相关的输出，包括表和图表。模型视图中的表包括焦点个案的 *k* 个最近邻元素和距离，分类响应变量的分类以及误差摘要。模型视图中的图形输出包括选择误差日志、特征重要性图表、特征空间图表、对等图表和象限图。请参阅 第 79 页的『模型视图』主题以获取更多信息。

文件

- **将模型导出到 XML**。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。如果已经指定拆分文件，此选项不可用。
- **导出焦点个案和 k 个最近邻元素之间的距离**。对于每个焦点个案，为其 *k* 个最近邻元素（来自训练样本）和相应的 *k* 个最近距离创建单独的变量。

选项

用户缺失值。要在分析中包含个案，分类变量必须具有有效值。通过这些控制可以决定是否将用户缺失值在分类变量中视为有效值。

系统缺失值和刻度变量缺失值总是被视为无效。

模型视图

在“输出”选项卡中选择**图表和表**时，过程会在查看器中创建“最近邻元素模型”对象。激活（双击）该对象，可获得模型的交互式视图。此模型视图有 2 个面板窗口:

- 第一个面板显示模型概览，称为主视图。
- 第二个面板显示两种视图类型之一:

  辅助模型视图显示有关模型的更多信息，但并不专注于模型本身。

  当用户深入查看主视图某个部分时，链接视图显示有关某个模型特征的详细信息。

缺省情况下，第一个面板显示特征空间，第二个面板显示变量重要性图表。如果变量重要性图表不可用，即“特征”选项卡上未选中**按重要性加重指定权重特征**，此时显示“视图”下拉列表中的第一个可用的视图。

如果视图没有可用信息，那么禁用其在“视图”下拉列表中的项文本。

特征空间

特征空间图表是有关特征空间（如果存在 3 个以上特征，那么为子空间）的交互式图形。每条轴代表模型中的某个特征，图表中的点位置显示个案这些特征在训练和坚持分区中的值。

关键字。除了特征值外，图中的点还传递其它信息。

- 其形状表示点所属的分区，即训练或坚持分区。
- 点的颜色/阴影表示该个案的目标值，不同的颜色值等于分类目标的类别，阴影则表示连续目标的值范围。训练分区的指示值为观察值；对于坚持分区，那么为预测值。如果未指定目标，那么不会显示此键。
- 较粗的概要表示个案为焦点个案。显示的焦点个案链接到它们的 *k* 个最近邻元素。

控制和互动。使用图表中的一些控制可以探索特征空间。

- 可以选择在图表中显示哪个特征子集，还可更改在维度上表示哪些特征。
- “焦点个案”仅仅是在“特征空间”图表中选择的点。如果指定了焦点个案变量，那么初始情况下会选中代表焦点个案的点。不过,任何点都可以暂时成为焦点个案，只要您将其选中。可以使用用于选择点的“常规”控件，即，单击一个点将选中该点并取消选中所有其他点；按下 Ctrl 键并单击一个点会将其添加到选择的点集合。链接的视图，如对等图表，将根据在“特征空间”中选择的个案自动更新。
- 您可以更改为焦点个案显示的最近邻元素数目 (*k*)。
- 在图表中的点上方悬停，可以显示工具提示以及个案标签值，或个案编号（如果未定义个案标签），以及观察和预测目标值。
- 使用“重置”按钮可以将“特征空间”恢复为其原始状态。

添加和删除字段/变量

您可以添加新字段/变量到特征空间或删除当前显示的字段/变量。

变量调色板

必须在添加和删除变量之前显示“变量”调色板。要显示“变量”调色板，“模型查看器”必须在“编辑”方式，同时必须在特征空间中选择了一个个案。

1. 要将“模型查看器”设置为“编辑”方式，请从菜单中选择:

   视图 > 编辑方式

2. 进入“编辑”方式后，单击特征空间中的任意个案。
3. 要显示“变量”调色板，请从菜单中选择:

   查看 > 选用板 > 变量

   “变量”调色板在特征空间中列出所有变量。变量名称旁边的图标表示变量的测量级别。
4. 要暂时更改变量的测量级别，右键单击变量调色板中的变量，然后选择一个选项。

变量区域

变量被添加到特征空间中的“区域”里。要显示区域，从“变量”调色板拖动变量或选择**显示区域**。

特征空间有 *x*、*y* 和 *z* 轴的区域。

移动“变量”到“区域”中

以下是移动变量到区域中的一些一般规则和技巧:

- 要移动变量到区域中，从“变量”调色板单击并拖动变量，然后放到区域中。如果您选择**显示区域**，您还可以右键单击一个区域并选择您希望添加到区域中的变量。
- 如果您从“变量”调色板拖动变量到一个已经被另一个变量占用的区域，旧的变量将被新的变量代替。
- 如果您从一个区域拖动变量到另一个已经被另一个变量占用的区域，这两个变量将交换位置。
- 单击区域中的 X 从该区域删除变量。
- 如果在可视化中有多个图形元素，每个图形元素可以拥有其自己的关联变量区域。首先，选择图形元素。

变量重要性

通常，您将需要将建模工作专注于最重要的变量，并考虑删除或忽略那些最不重要的变量。变量重要性图表可以在模型估计中指示每个变量的相对重要性，从而帮助您实现这一点。由于它们是相对值，因此显示的所有变量的值总和为 1.0。变量重要性与模型精度无关。它只与每个变量在预测中的重要性有关，而不涉及预测是否精确。

对等

该图表显示焦点个案及其在每个特征和目标上 *k* 个最近邻元素。它仅在“特征空间”图表中选择了焦点个案时可用。

链接行为。对等图表以两种方式链接到“特征空间”。

- 在特征空间中所选的个案（焦点个案）显示在对等图表中，也包括其 *k* 个最近邻元素。
- 在对等图表中使用在特征空间中所选的 *k* 值。

最近邻元素距离

该表只显示焦点个案的 *k* 个最近邻元素与距离。它仅当在“变量”选项卡上指定了焦点个案标识时可用，且仅显示由此变量标识的焦点个案。

每行:

- **焦点个案**列包含焦点个案的个案标签变量值；如果未定义个案标签，那么此列包含焦点个案的个案编号。
- “最近邻元素”组下方的第 *i* 列包含焦点个案的第 *i* 个最近邻元素的个案标签变量值；如果未定义个案标签，那么此列包含焦点个案的第 *i* 个最近邻元素的个案编号。
- “最近距离”组下方的第 *i* 列包含第 *i* 个最近邻元素到焦点个案的距离。

象限图

该图表显示焦点个案及其在散点图（点图，取决于目标的测量级别）上 *k* 个最近邻元素。目标在 *y* 轴上，刻度特征在 *x* 轴上，按特征划分面板。它仅当存在目标，且在“特征空间”图表中选择了焦点个案时可用。

- 在训练分区的变量平均值处，为连续变量绘制了参考线。

特征选择误差日志

对于该图表上的点，其 *y* 轴值为模型的误差（误差率或平方和误差，取决于目标的测量级别），*x* 轴上列出模型的特征（加上 *x* 轴左侧的所有特征）。该图表仅当存在目标，且特征选择有效时可用。

K 选择误差日志

对于该图表上的点，其 *y* 轴值为模型的误差（误差率或平方和误差，取决于目标的测量级别），*x* 轴上为最近邻元素数目（*k*)。该图表仅当存在目标，且 *k* 选择有效时可用。

k 和特征选择误差日志

这些是特征选择图表（请参阅『特征选择误差日志』），并按 *k* 划分面板。仅当存在目标，且 *k* 和特征选择均有效时，此图表才可用。

分类表

该表显示按分区对目标观察与预测值的交叉分类。它仅当存在分类目标时可用。

- 坚持分区中的（**缺失**）行包含在目标上具有缺失值的坚持个案。这些个案对“坚持样本：整体百分比”值有贡献，但对“正确百分比”值无影响。

误差摘要

该表仅当存在目标变量时可用。它显示模型的相关误差；即，连续目标的平方和误差以及分类目标的误差率（100% - 总体正确百分比）。

第 21 章 判别分析

判别分析为组成员身份构建预测模型。该模型将基于可提供组间最佳区分的预测变量的线性组合，包含判别函数（或，对两个以上的组，包含一组判别函数）。这些函数根据组成员身份已知的个案样本生成；然后，可以将这些函数应用于具有预测变量测量值，但具有未知组成员身份的新个案。

注：分组变量可有两个以上的值。分组变量的代码必须为整数，但是，您需要指定其最小值和最大值。将从分析中排除具有位于边界以外的值的个案。

示例。平均而言，温带国家/地区的人比热带的人每天消耗的卡路里多，且温带地区城市居民的比例比热带大。一名研究员想将这些信息组合为函数，以确定通过某个人区分出这两组国家/地区的有效程度。该研究员认为人口数量和经济状况信息可能也很重要。判别分析允许您估计线性判别函数的系数，线性判别函数看起来像多重线性回归方程式的右侧部分。即，使用系数 *a*、*b*、*c* 和 *d*，函数为:

```
D = a * climate + b * urban + c * population + d * gross domestic product per capita
```

如果这些变量可用于辨别这两个气候带，温带国家/地区和热带国家/地区将有不同的 *D* 值。如果您使用逐步式变量选择法，您可能发现不需要在函数中包含所有四个变量。

统计。对于每个变量：平均值、标准差和单变量 ANOVA。对于每项分析：Box *M*、组内相关性矩阵、组内协方差矩阵、分组协方差矩阵以及总体协方差矩阵。对于每个典型判别函数：特征值、方差百分比、典型相关性、Wilks Lambda 和卡方。对于每个步骤：先验概率、Fisher 函数系数、非标准化函数系数以及每个典型函数的 Wilks Lambda。

判别分析数据注意事项

数据。分组变量必须含有有限数目的不同类别，且编码为整数。名义自变量必须被重新编码为哑元变量或对比变量。

假设。个案应为独立的。预测变量应有多变量正态分布，组内方差-协方差矩阵在组中应等同。组成员身份假设为互斥的（即，不存在属于多个组的个案），且全体为穷举的（即，所有个案均是组成员）。组成员身份为真正的分类变量时，此过程最有效；如果组成员身份基于连续变量的值（例如，高智商与低智商），那么请考虑使用线性回归以利用由连续变量本身提供的更为丰富的信息。

获得判别分析

1. 从菜单中选择:

   分析 > 分类 > 判别...
2. 选择一个整数值的分组变量并单击**定义范围**以指定感兴趣的类别。
3. 选择自变量（预测变量）。（如果分组变量不含整数值，那么“转换”菜单中的“自动重新编码”将创建含整数值的变量。）
4. 选择输入自变量的方法。
   - **一起输入自变量**。同时输入所有满足容差标准的自变量。
   - **使用步进法**。使用逐步式分析控制变量输入与移去。
5. 根据需要，选择含选择变量的个案。

© Copyright IBM Corporation 1989, 2014

判别分析: 定义范围

指定用于分析的分组变量的最小值和最大值。在判别分析中不使用值在该范围外的个案，但将基于分析结果将这些个案划分到某个现有组中。最小值和最大值必须为整数。

判别分析: 选择个案

选择用于分析的个案:

1. 在“判别分析”对话框中，选择一个选择变量。
2. 单击**值**以输入整数作为选择值。

仅使用选择变量具有指定值的个案来推导判别函数。同时为选定的和未选定的个案生成统计和分类结果。此过程提供一种机制，通过这种机制可以基于以前存在的数据对新个案进行分类，或将您的数据划分成训练子集和检验子集，以执行对所生成模型的验证。

判别分析: 统计

描述性。可用选项为平均值（包括标准差）、单变量 ANOVA 以及 Box 的 *M* 检验。

- *平均值 (Means).* 显示自变量的总平均值、组平均值和标准差。
- *单变量 ANOVA (Univariate ANOVAs).* 为每个自变量的组平均值的等同性执行单向方差检验分析。
- *Box M*。组协方差矩阵的等同性检验。对于足够大的样本，不显著的 p 值表示断定矩阵不同的证据不足。该检验对于偏离多变量正态性很敏感。

函数系数。可用的选项有 Fisher 的分类系数和未标准化的系数。

- *Fisher's.* 显示可以直接用于分类的 Fisher 分类函数系数。为每个组获得一组单独的分类函数系数，将一个个案分配给该组，该个案对此组具有最大判别分数（分类函数值）。
- *未标准化 (Unstandardized).* 显示未标准化的判别函数系数。

矩阵。可用的自变量系数矩阵有组内相关性矩阵、组内协方差矩阵、分组协方差矩阵和总体协方差矩阵。

- *组内相关性*。显示汇聚的组内相关性矩阵，获取该矩阵的方法是在计算相关性之前，求得所有组的单个协方差矩阵的平均值。
- *组内协方差*。显示汇聚的组内协方差矩阵，该矩阵与总协方差矩阵可能不同。获取该矩阵的方法是，求得所有组的单个协方差矩阵的平均值。
- *分组协方差*。显示每个组的分离协方差矩阵。
- *总体协方差*。显示来自所有个案的协方差矩阵，就好像它们来自一个样本一样。

判别分析: 步进法

方法。选择用于输入或移去新变量的统计。可用统计有 Wilks 的 lambda、未解释的方差、马氏距离、最小 *F* 比以及 Rao 的 *V*。使用 Rao 的 *V*，您可为要输入的变量指定在 *V* 中增加的最小值。

- *Wilks Lambda*。一种用于逐步判别分析的变量选择方法，它基于变量能在多大程度上降低 Wilks 的 lambda 来选择要输入到方程中的变量。在每一步，均是输入能使总体 Wilks 的 lambda 最小的变量。
- *未解释方差*。在每一步，均是输入能使组间未解释变动合计最小的变量。
- *马氏距离*。自变量上个案的值与所有个案的平均值相异程度的测量。大的马氏距离表示个案在一个或多个自变量上具有极值。
- *最小 F 比*。一种逐步分析中的变量选择方法，它基于使从组间马氏距离计算得到的 F 比最大。

- *Rao V*。组平均值之间的差分的测量。也称为 Lawley-Hotelling 轨迹。在每一步，能使 Rao 的 V 增加最大的变量被选进来。选择此选项之后，请输入要进入分析，变量必须具有的最小值。

标准。可用的备用项包括**使用 F 值**和**使用 F 的概率**。请输入进入变量和移去变量的值。

- *使用 F 值*。如果变量的 F 值大于“输入”值，那么该变量输入模型，如果 F 值小于“剔除”值，那么该变量从模型中移去。“输入”值必须大于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请降低“输入”值。要将更多的变量从模型中移去，请增大“剔除”值。
- *使用 F 的概率*。如果变量的 F 值的显著性水平小于“输入”值，那么将该变量选入到模型中，如果该显著性水平大于“剔除”值，那么将该变量从模型中移去。“输入”值必须小于“剔除”值，且两者均必须为正数。要将更多的变量选入到模型中，请增加“输入”值。要将更多的变量从模型中移去，请降低“剔除”值。

输出。**步进摘要**显示完成每一步后所有变量的统计；**两两组间距离的 F 值**显示每一组对的成对 *F* 比的矩阵。

判别分析: 分类

先验概率。此选项确定对于组成员身份的先验知识，是否调整分类系数。

- **所有组相等**。假设所有组的先验概率相等；这对系数没有影响。
- **根据组大小计算**。样本中的观察组大小决定组成员身份的先验概率。例如，如果分析中包括的 50% 的观察值属于第一组，25% 属于第二组，25% 属于第三组，那么会调整分类系数以增加第一组相对于其他两组的成员身份可能性。

输出。可用的显示选项有个案结果、摘要表和留一分类。

- *个案结果*。为每个个案显示实际组的代码、预测组、后验概率和判别分数。
- *摘要表*。基于判别分析，正确地和不正确地指定给每个组的个案数。有时称为“混乱矩阵”。
- *留一分类*。分析中的每个个案由除该个案之外的所有个案生成的函数来进行分类。这也称为“U 方法”。

使用平均值替换缺失值。选择该选项，仅在分类阶段用自变量的平均值代替缺失值。

使用协方差矩阵。您可用选择使用组内协方差矩阵或分组协方差矩阵对个案进行分类。

- *在组内*。汇聚的组内协方差矩阵用来对个案分类。
- *分组*。分组协方差矩阵用于分类。由于分类基于判别函数（而非基于原始变量），因此该选项并不总是等同于二次判别。

图。可用的图选项有合并组、分组和区域图。

- *合并组*。创建前两个判别函数值的所有组散点图。如果只有一个函数，那么转而显示一个直方图。
- *分组*。创建前两个判别函数值的分组散点图。如果只有一个函数，那么转而显示直方图。
- *区域图*。用于基于函数值将个案分类到组的边界图。其个数对应于个案分类到的组数。每个组的平均值在其边界内用一个星号表示。如果只有一个判别函数，那么该图不会显示。

判别分析: 保存

您可以向活动数据文件添加新变量。可用的选项有预测的组成员身份（单个变量）、判别分数（解中每个判别函数均有一个变量）和已给出判别分数的组成员身份的概率（每一组有一个变量）。

您还可以将模型信息导出到指定的 XML 格式文件中。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。

DISCRIMINANT 命令的附加功能

使用命令语法语言还可以:

- 执行多个判别分析（使用一个命令），并控制输入变量的顺序（使用 ANALYSIS 子命令）。
- 指定分类的先验概率（使用 PRIORS 子命令）。
- 显示旋转的模式和结构矩阵（使用 ROTATE 子命令）。
- 限制提取的判别函数的数量（使用 FUNCTIONS 子命令）。
- 将分类限于为分析而选择（或未选择）的个案（使用 SELECT 子命令）。
- 读取并分析相关性矩阵（使用 MATRIX 子命令）。
- 写相关性矩阵以用于以后的分析（使用 MATRIX 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 22 章 因子分析

因子分析尝试识别出基础变量（或称**因子**）来解释在一组观察到的变量中体现的相关模式。因子分析通常用于数据降维，其目的是识别出少数几个因子来解释大多数在众多显性变量中所观测到的方差。因子分析也可用于生成关于因果机制的假设或过滤变量以用于随后的分析（例如：在执行线性回归分析之前识别共线性）。

因子分析过程提供了高度的灵活性:

- 有 7 种因子抽取的方法。
- 有 5 种旋转方法，包括直接 Oblimin 方法和非正交旋转的最优斜交。
- 有 3 种计算因子得分的方法，并且得分可以另存为变量以进行进一步分析。

示例。什么基础态度使人们回答政治调查上的问题？检查调查项中的相关性显示，项的各种子组有显著的交迭 – 关于税的问题显得彼此相关，关于军事的问题显得彼此相关，等等。使用因子分析，您可以调查基础因子的数量，并且，在许多情况下，还可以识别这些因子在概念上所代表的含义。此外，您可以计算每个响应者的因子得分，然后这些得分可以用于以后的分析。例如，您可以建立 logistic 回归模型以根据因子得分预测投票行为。

统计。对于每个变量：有效个案数、平均值和标准差。对于每项因子分析：变量的相关性矩阵，包括显著性水平、行列式和逆；再生相关性矩阵，包括反映像；初始解（公因子方差、特征值以及解释的方差所占的百分比）；对取样充分性的 Kaiser-Meyer-Olkin 度量以及 Bartlett 球形度检验；未旋转的解，包括因子载荷、公因子方差和特征值；以及旋转解，包括旋转的模式矩阵和转换矩阵。对于斜交旋转：旋转的模式和结构矩阵；因子得分系数矩阵和因子协方差矩阵。图：特征值的碎石图以及前两个或前三个因子的载荷图。

因子分析数据注意事项

数据。变量在*区间*或*比率*级别应该是定量变量。分类数据（例如：宗教或原产国家/地区）不适合因子分析。可计算 Pearson 相关性系数的数据应该适合于因子分析。

假设。对于每对变量，数据应具有二元正态分布，且观察值应是独立的。因子分析模型指定变量是由公共因子（模型估计的因子）和特殊因子（不在观察到的变量之间交迭）确定的；计算的估计值所基于的假设是所有唯一因子相互之间不相关并与公共因子不相关。

获取因子分析

1. 从菜单中选择:

   分析 > 降维 > 因子...
2. 选择用于因子分析的变量。

因子分析: 选择个案

选择用于分析的个案:

1. 选择一个选择变量。
2. 单击**值**以输入整数作为选择值。

因子分析中仅使用具有该选择变量值的个案。

© Copyright IBM Corporation 1989, 2014

因子分析: 描述

统计。 **一元描述**包括每个变量的平均值、标准差和有效个案数。**初始解**显示初始公因子方差、特征值和已解释方差的百分比。

相关性矩阵。可用选项为系数、显著性水平、行列式、KMO 和 Bartlett 球形度检验、逆、再生和反映像。

- *KMO 和 Bartlett 球形度检验。*取样足够度的 Kaiser-Meyer-Olkin 度量检验变量之间的偏相关性是否较小。Bartlett 的球形度检验可检验相关性矩阵是否为恒等矩阵，该检验可以指示因子模型不适当。
- *再生。*从因子解估计的相关性矩阵。还显示残差（估计相关性和观察相关性之间的差分）。
- *反映象。*反映像相关性矩阵包含偏相关系数的相反数，而反映像协方差矩阵包含偏协方差的相反数。在一个好的因子模型中，大部分非对角线的元素将会很小。变量的取样充分性度量显示在反映像相关性矩阵的对角线上。

因子分析: 抽取

方法。使您可以指定因子抽取的方法。可用方法为主成分分析、未加权最小二次方、广义最小二乘法、极大似然、主轴因子分解、Alpha 因子分解和映像因子分解。

- *主成分分析 (Principal Components Analysis).* 一种因子抽取方法，用于形成观察变量的不相关的线性组合。第一个成分具有最大的方差。后面的成分对方差的解释的比例逐渐变小，它们相互之间均不相关。主成分分析用来获取最初因子解。它可以在相关性矩阵是奇异矩阵时使用。
- *未加权最小二次方法 (Unweighted Least-Squares Method).* 一种因子抽取方法，该方法可以使观察的相关性矩阵和再生的相关性矩阵之间的差的平方值之和最小（忽略对角线）。
- *广义最小二乘法 (Generalized Least-Squares Method).* 一种因子抽取方法，该方法可以使观察的相关性矩阵和再生的相关性矩阵之间的差的平方值之和最小。相关系数要进行加权。权重为他们单值的倒数，这样单值高的变量，其权重比单值低的变量的权重小。
- *极大似然法 (Maximum-Likelihood Method).* 一种因子抽取方法，在样本来自多变量正态分布的情况下，它生成的参数估计最有可能生成了观察到的相关性矩阵。将变量单值的倒数作为权重对相关性进行加权，并使用迭代算法。
- *主轴因子分解 (Principal Axis Factoring).* 一种从初始相关性矩阵抽取因子的方法，在初始相关性矩阵中，多元相关系数的平方放置于对角线上作为公因子方差的初始估计值。这些因子载荷用来估计替换对角线中的旧公因子方差估计值的新的公因子方差。继续迭代，直到某次迭代和下次迭代之间公因子方差的改变幅度能满足抽取的收敛性条件。
- *α.* 一种因子抽取方法，它将分析中的变量视为来自潜在变量全体的一个样本。此方法使因子的 alpha 可靠性最大。
- *映像因子分解 (Image Factoring).* 由 Guttman 开发的因子抽取方法，它基于映像理论。变量的公共部分（称为偏映像）定义为其对剩余变量的线性回归，而非假设因子的函数。

分析。使您可以指定相关性矩阵或协方差矩阵。

- **相关性矩阵。**在分析中使用不同的刻度测量变量时很有用。
- **协方差矩阵。**当您想将因子分析应用于每个变量具有不同方差的多个组时很有用。

抽取。可以保留特征值超过指定值的所有因子，也可以保留特定数量的因子。

输出。使您可以请求未旋转的因子解和特征值的碎石图。

- *未旋转的因子解 (Unrotated Factor Solution).* 显示未旋转的因子载荷（因子模式矩阵）、公因子方差和因子解的特征值。

- *碎石图*。与每个因子相关联的方差的图。该图用于确定应保持的因子个数。通常该图显示大因子的陡峭斜率和剩余因子平缓的尾部之间明显的中断（碎石）。

最大收敛性迭代次数。使您可以指定算法估计解的过程所采取的最大步骤数。

因子分析: 旋转

方法。使您可以选择因子旋转的方法。可用的方法有最大方差、直接 Oblimin、最大四次方值、最大平衡值或最优斜交。

- *最大方差法 (Varimax Method)*. 一种正交旋转方法，它使得对每个因子有高负载的变量的数目达到最小。该方法简化了因子的解释。
- *直接 Oblimin 方法*。一种斜交（非正交）旋转方法。当 delta 等于 0（缺省值）时，解是最斜交的。delta 负得越厉害，因子的斜交度越低。要覆盖缺省的 delta 值 0，请输入小于等于 0.8 的数。
- *最大四次方值法 (Quartimax Method)*. 一种旋转方法，它可使得解释每个变量所需的因子最少。该方法简化了观察到的变量的解释。
- *最大平衡值法 (Equamax Method)*. 一种旋转方法，它是简化因子的最大方差法与简化变量的最大四次方值法的组合。它可以使得高度依赖因子的变量的个数以及解释变量所需的因子的个数最少。
- *最优斜交旋转 (Promax Rotation)*. 斜交旋转，可使因子相关联。该旋转可比直接最小斜交旋转更快地计算出来，因此适用于大型数据集。

输出。使您可以在旋转解上包含输出以及前两个或前三个因子的载荷图。

- *旋转解 (Rotated Solution)*. 必须选择旋转方法才能获得旋转解。对于正交旋转，会显示已旋转的模式矩阵和因子转换矩阵。对于斜交旋转，会显示模式、结构和因子相关性矩阵。
- *载荷图 (Factor Loading Plot)*. 前三个因子的三维因子载荷图。对于双因子解，那么显示二维图。如果只抽取了一个因子，那么不显示图。如果要求旋转，那么图会显示旋转解。

最大收敛性迭代次数。使您可以指定算法执行旋转所采取的最大步骤数。

因子分析: 得分

保存为变量。为最终解中的每个因子创建一个新变量。

方法。计算因子得分的可选方法有回归、Bartlett 和 Anderson-Rubin。

- *回归法 (Regression Method)*. 一种估计因子得分系数的方法。生成的分数的平均值为 0，方差等于估计的因子分数和真正的因子值之间的平方多相关性。即使因子是正交的，分数也可能相关。
- *Bartlett 得分*。一种估计因子得分系数的方法。所产生分数的平均值为 0。使整个变量范围中所有唯一因子的平方和达到最小。
- *Anderson-Rubin 方法 (Anderson-Rubin Method)*. 一种估计因子得分系数的方；它对 Bartlett 方法做了修正，从而确保被估计的因子的正交性。生成的分数平均值为 0，标准差为 1，且不相关。

显示因子得分系数矩阵。显示与变量相乘以获取因子得分的系数。还显示因子得分之间的相关性。

因子分析: 选项

缺失值。允许您指定如何处理缺失值。可用选项为*按列表*排除个案，*成对*排除个案，或替换为平均值。

系数显示格式。使您可以控制输出矩阵的各个方面。按大小对系数进行排序，并排除绝对值小于指定值的系数。

FACTOR 命令的附加功能

使用命令语法语言还可以:

- 在抽取和旋转过程中指定迭代的收敛性标准。
- 指定单独的旋转因子图。
- 指定保存多少因子得分。
- 指定用于主轴因子分解方法的对角线值。
- 将相关性矩阵或因子载荷矩阵写入磁盘以便以后分析。
- 读取和分析相关性矩阵或因子载荷矩阵。

请参阅*命令语法参考*以获取完整的语法信息。

第 23 章 选择聚类过程

可以使用二阶、系统或 K 平均值聚类分析过程来执行聚类分析。每个过程使用不同的算法来创建聚类，并且每个过程所具有的选项在其他过程中不可用。

二阶聚类分析。对很多应用而言，二阶聚类分析过程是首选的方法。它提供以下独特的功能:

- 除了用于在聚类模型之间进行选择的测量之外，还可自动选择最佳聚类数目。
- 能够同时根据分类和连续变量创建聚类模型。
- 能够将聚类模型保存到外部 XML 文件，然后读取该文件并使用较新的数据来更新聚类模型。

此外，“二阶聚类分析”过程可以分析大数据文件。

系统聚类分析。系统聚类分析过程只限于较小的数据文件（要聚类的对象只有数百个），但具有以下独特功能:

- 能够对个案或变量进行聚类。
- 能够计算可能解的范围，并为其中的每一个解保存聚类成员。
- 有多种方法可用于聚类形成、变量转换以及度量各聚类之间的非相似性。

只要所有变量的类型相同，“系统聚类分析”过程就可以分析区间（连续）、计数或二值变量。

K 平均值聚类分析。K 平均值聚类分析过程只限于连续数据，要求预先指定聚类数目，但它具有以下独特的功能:

- 能够保存每个对象与聚类中心之间的距离。
- 能够从外部 IBM SPSS Statistics 文件中读取初始聚类中心，并将最终的聚类中心保存到该文件中。

此外，K 平均值聚类分析过程可以分析大数据文件。

© Copyright IBM Corporation 1989, 2014

第 24 章 二阶聚类分析

“二阶聚类分析”过程是一个探索工具，用来揭示数据集中的自然分组（或聚类），如果不揭示，这些分组是不明显的。此过程使用的算法有多个不错的特征使其区分于传统聚类技术:

- **分类变量和连续变量的处理**。通过假设变量是独立的，可以假设分类变量和连续变量服从联合多项正态分布。
- **聚类数的自动选择**。通过跨不同的聚类解比较模型选择准则的值，此过程可以自动确定最优的聚类数。
- **可缩放性**。通过构造摘要记录的聚类特征 (CF) 树，二阶算法允许您分析大型数据文件。

示例。零售和消费者产品公司定期地对描述客户的购买习惯、性别、年龄、收入水平等的数据应用聚类技术。这些公司为每个消费者群体设计营销和产品开发战略，以增加销售额和建立品牌忠诚度。

距离测量。此选项确定如何计算两个聚类之间的相似性。

- **对数相似性**。该似然度量假设变量服从某种概率分布。假设连续变量是正态分布，而假设分类变量是多项分布。假设所有变量均是独立的。
- **欧几里德距离**。欧几里德距离测量是两个聚类之间的“直线”距离。它只能用于所有变量连续的情况。

聚类数。此选项允许您指定如何确定聚类数。

- **自动确定**。此过程将使用在“聚类准则”组中指定的准则，自动确定“最好”的聚类数。或者，还可以输入一个正整数指定过程应考虑的最大聚类数。
- **指定固定值**。允许您固定解中的聚类数。输入正整数。

连续变量计数。此组提供了在“选项”对话框中指定的连续变量标准化的摘要。请参阅 第 94 页的『二阶聚类分析: 选项』主题以获取更多信息。

聚类准则。此选项确定自动聚类算法如何确定聚类数。可以指定 Bayesian 信息标准 (BIC) 或 Akaike 信息标准 (AIC)。

二阶聚类分析数据注意事项

数据。此过程既处理连续变量也处理分类变量。个案代表要聚类的对象，变量代表聚类所基于的属性。

个案顺序。注意，聚类特征树和最终解可能取决于个案顺序。要使顺序的影响降至最低程度，可随机个案等级排序的顺序。您可能想要通过以不同随机顺序排序的案例来得到多个不同的解，以验证给定解的稳定性。如果由于文件非常大而无法获取多个不同的解，可使用以不同的随机顺序排序的个案样本运行多次。

假设。似然距离测量假设聚类模型中的变量是独立的。而且，假设每个连续变量具有正态（高斯）分布，假设每个分类变量具有多项分布。经验内部检验表明，此过程对于违反独立性假设和分布假设均相当稳健，但您应尝试了解这些假设符合的程度。

使用双变量相关性过程可以检验两个连续变量的独立性。使用交叉表过程可检验两个分类变量的独立性。使用平均值过程可以检验连续变量与分类变量之间的独立性。使用探索过程可以检验连续变量的正态性。使用卡方检验过程可以检验分类变量是否具有指定的多项分布。

获取二阶聚类分析

1. 从菜单中选择:

© Copyright IBM Corporation 1989, 2014

分析 > 分类 > 二阶聚类...

2. 选择一个或多个分类变量或连续变量。

根据需要，您可以:

- 调整构造聚类的标准。
- 选择噪声处理、内存分配、变量标准化和聚类模型输入的设置。
- 请求模型查看器输出。
- 将模型结果保存到工作文件或外部 XML 文件。

二阶聚类分析: 选项

离群值处理。 该组允许您在聚类特征 (CF) 树填满的情况下，在聚类过程中特别地处理离群值。如果 CF 树的叶节点中不能接受更多的个案，且所有叶节点均不能拆分，那么说明 CF 树已满。

- 如果选择噪声处理且 CF 树填满，那么在将稀疏叶子中的个案放到“噪声”叶子中后，树将重新生长。如果某个叶子包含的个案数占最大叶大小的百分比小于指定的百分比，那么将该叶子视为稀疏的。树重新生长之后，如有可能，离群值将放置在 CF 树中。否则，将放弃离群值。
- 如果不选择噪声处理且 CF 树填满，那么它将使用较大的距离更改阈值来重新生长。最终聚类之后，不能分配到聚类的变量标记为离群值。离群值聚类被赋予标识号 -1，并且不会包括在聚类数的计数中。

内存分配。此组允许您以兆字节 (MB) 为单位，指定聚类算法应使用的最大的内存量。如果此过程超过了此最大值，那么将使用磁盘存储内存中放不下的信息。请指定大于等于 4 的数。

- 请咨询系统管理员以获取您可以在系统上指定的最大值。
- 如果此值太小，那么算法可能无法找到正确或指定数目的聚类。

变量标准化。聚类算法处理标准化连续变量。任何未标准化的连续变量都应保留为“要标准化的变量”列表中的变量。为了节省部分时间和计算工作，您可以选择任何已标准化的连续变量作为“假定已标准化的变量”列表中的变量。

高级选项

CF 树调节准则。以下聚类算法设置特别地应用到聚类特征 (CF) 树，且应谨慎地更改:

- **初始距离更改阈值**。这是用来使 CF 树生长的初始阈值。如果将给定的个案插入到 CF 树的叶子中将生成小于阈值的紧度，那么不会拆分叶子。如果紧度超过阈值，那么会拆分叶子。
- **最大分支（每个叶节点）**。叶节点可以具有的最大子节点数。
- **最大树深度**。CF 树可以具有的最大级别数。
- **可能的最大节点数**。此值指示过程可能生成的最大 CF 树节点数，这基于函数 $(b^{d+1} - 1) / (b - 1)$，其中 b 是最大分支数，d 是最大树深度。请注意，非常大的 CF 树可能会耗尽系统资源，从而对过程的性能产生不利影响。每个节点最少需要 16 个字节。

聚类模型更新。此组允许您导入和更新在先前分析中生成的聚类模型。输入文件以 XML 格式包含 CF 树。然后将使用活动文件中的数据更新模型。必须在主对话框中以与先前分析中指定的顺序相同的顺序选择变量名。除非您专门将新的模型信息写到相同的文件名中，否则该 XML 文件保持不变。请参阅第 95 页的『二阶聚类分析: 输出』主题以获取更多信息。

如果指定聚类模型更新，那么使用与为原始模型指定的 CF 树的生成相关的选项。具体而言，会使用已保存模型的距离测量、噪声处理、内存分配或 CF 树调节准则设置，将忽略对话框中这些选项的任何设置。

注: 执行聚类模型更新时，此过程假设不使用活动数据集中任何选定的个案创建原始聚类模型。此过程还假设用在模型更新中的个案与用于创建原始模型的个案来自同一总体；也就是说，假设连续变量的平均值和方差以及分类变量的级别在两个个案组上相同。如果“新的”和“旧的”个案组来自不同的总体，那么应对组合个案组运行“二阶聚类分析”过程以获取最佳结果。

二阶聚类分析: 输出

输出。该组提供显示聚类结果的选项。

- **透视表**。结果将显示在透视表中。
- **图表和表（在模型查看器中）**。结果将显示在模型查看器中。
- **评估字段**。这可为未在聚类创建中使用的变量计算聚类数据。通过在“显示”子对话框中选择评估字段，可以在模型查看器中将其与输入特征一起显示。带有缺失值的字段将被忽略。

工作数据文件。该组允许您将变量保存到活动数据集。

- **创建聚类成员变量**。此变量包含每个个案的聚类标识号。此变量的名称为 *tsc_n*，其中 *n* 是一个正整数，表示在给定会话中由此过程完成的活动数据集保存操作的序数。

XML 文件。最终聚类模型和 CF 树是两类可以以 XML 格式导出的输出文件。

- **导出最终模型**。最终聚类模型以 XML (PMML) 格式导出到指定文件中。您可以使用该模型文件以应用模型信息到其他数据文件用于评分目的。
- **导出 CF 树**。此选项允许您保存聚类树的当前状态，并在以后使用较新的数据对其进行更新。

聚类查看器

聚类模型通常用于根据所检查变量查找具有类似记录的组（聚类），其中同组成员间的相似性高而不同组成员间的相似性低。结果可用于识别原本不明显的关联。例如，通过对客户偏好、收入水平和购物习惯的聚类分析，可以识别出对某种市场营销活动更可能做出反应的客户类型。

有两种方法可以解释聚类显示中的结果:

- 检查聚类以确定该聚类的唯一特征。*是否有一个聚类包含所有高收入借款人？此聚类是否包含比其他聚类更多的记录？*
- 检查各聚类上的字段以确定值在聚类间的分布情况。*个人的教育水平是否决定其在聚类中的成员资格？高信用得分是否在一个聚类或另一个聚类的成员资格之间加以区分？*

使用“聚类查看器”中的主视图和各个链接视图，可以清楚回答这些问题。

要查看有关聚类模型的信息，激活（双击）“查看器”中的“模型查看器”对象。

聚类查看器

“聚类查看器”包含两个面板，主视图位于左侧，链接或辅助视图位于右侧。有两个主视图:

- 模型摘要（缺省视图）。请参阅 第 96 页的『模型摘要视图』主题以获取更多信息。
- 聚类。请参阅 第 96 页的『聚类视图』主题以获取更多信息。

有四个链接/辅助视图:

- 预测变量的重要性.请参阅 第 97 页的『聚类预测变量重要性视图』主题以获取更多信息。
- 聚类大小（缺省视图）。请参阅 第 97 页的『聚类大小视图』主题以获取更多信息。
- 单元格分布。请参阅 第 97 页的『单元格分布视图』主题以获取更多信息。

- 聚类比较。请参阅 第 98 页的『聚类比较视图』主题以获取更多信息。

模型摘要视图

“模型摘要”视图显示聚类模型的快照或摘要，包括加阴影以表示结果较差、尚可或良好的聚类结合和分离的 Silhouette 测量。该快照可让您快速检查质量是否较差，如果较差，您可返回建模节点修改聚类模型设置以生成较好的结果。

结果较差、尚可和良好是基于 Kaufman 和 Rousseeuw (1990) 关于聚类结构解释的研究成果来判定的。在“模型摘要”视图中，良好的结果表示数据将 Kaufman 和 Rousseeuw 的评级反映为聚类结构的合理迹象或强迹象，尚可的结果将其评级反映为弱迹象，而较差的结果将其评级反映为无明显迹象。

针对所有记录计算 (B-A) / max(A,B) 的平均值，其中 A 是记录与其聚类中心的距离，而 B 是记录与非所属最近聚类中心的距离。Silhouette 系数为 1 表示所有个案直接位于其聚类中心上。值为 -1 表示所有个案都位于另外某些聚类的聚类中心。值为 0 表示在正常情况下个案到其自身聚类中心与到最近其他聚类中心是等距的。

摘要所包含的表格具有以下信息:

- **算法**。所使用的聚类算法，例如“二阶”。
- **输入功能**。字段数量，也称为**输入**或**预测变量**。
- **聚类**。解中聚类的数量。

聚类视图

“聚类”视图包含一个聚类-特征网格，其中包括每个聚类的名称、大小和概要文件。

网格中的列包含以下信息:

- **聚类**。算法生成的聚类编号。
- **标签**。应用于每个聚类的任何标签（缺省为空白）。双击单元格输入描述聚类内容的标签，例如“豪华汽车买家”。
- **描述**。聚类内容的任何描述（缺省为空白）。双击单元格输入聚类描述；例如“年龄超过 55 岁、专业人员、收入超过 100,000 美元”。
- **大小**。每个聚类的大小，表示为总体聚类样本的百分比。网格中的每个大小单元格显示一个垂直条，其中显示聚类中的大小百分比、数值格式的大小百分比和聚类个案计数。
- **特征**。单个输入或预测变量，缺省按总体重要性排序。如果有列的大小相等，那么其以聚类编号的升序显示。

  总体特征重要性由单元格背景阴影的颜色表示；最重要的特征颜色最深；最不重要的特征则没有阴影。表格上方的向导指示与每个特征单元格颜色关联的重要性。

当鼠标悬停在单元格上时，会显示特征的全名/标签和单元格的重要性值。根据视图和特征类型，可能会显示其他信息。在“聚类中心”视图中，这包括单元格统计和单元格值；例如: “平均值: 4.32”。对于分类特征，单元格将显示最频繁（模态）类别的名称及其百分比。

在“聚类”视图中，您可以选择多种显示聚类信息的方式:

- 变换聚类和特征。请参阅 第 97 页的『变换聚类和特征』主题以获取更多信息。
- 排序特征。请参阅 第 97 页的『排序特征』主题以获取更多信息。
- 排序聚类。请参阅 第 97 页的『排序聚类』主题以获取更多信息。
- 选择单元格内容。请参阅 第 97 页的『单元格内容』主题以获取更多信息。

变换聚类和特征: 缺省情况下，聚类显示为列，特征显示为行。为翻转这种显示，单击**特征排序方式**按钮左侧的**变换聚类和特征**按钮。例如，当显示许多聚类时，您可能想要进行此操作，以减少查看数据所需的水平滚动量。

排序特征: **特征排序方式**按钮可使您选择特征单元格的显示方式:

- **总体重要性**。这是缺省的排序方式。特征以总体重要性的升序进行排序，排序方式在各聚类间相同。如果有特征具有同数重要性值，那么按照特征名称的升序列出同数特征。
- **聚类内重要性**。特征按照其相对于每个聚类的重要性进行排序。如果有特征具有同数重要性值，那么按照特征名称的升序列出同数特征。当选中此选项时，排序顺序通常因聚类而异。
- **名称**。特征按照名称的字母顺序进行排序。
- **数据顺序**。特征按照其在数据集中的顺序进行排序。

排序聚类: 缺省情况下，聚类按照大小的降序排序。**聚类排序方式**按钮可使您按照名称的字母顺序对其进行排序，或如果您创建了唯一标签，那么按照标签的字母顺序对其进行排序。

具有相同标签的特征按照聚类名称排序。如果聚类按照标签排序且您编辑了聚类的标签，那么自动更新排序顺序。

单元格内容: **单元格**按钮使您能够更改特征和评估字段的单元格内容的显示。

- **聚类中心**。缺省情况下，单元格显示特征名称/标签和每个聚类/特征组合的集中倾向。对于连续字段和具有分类字段的类别百分比的模式（最频繁出现的类别）显示平均值。
- **绝对分布**。显示特征名称/标签和每个聚类中特征的绝对分布。对于类别特征，显示条形图，其中叠放了按数据值的升序排序的类别。对于连续特征，显示平滑密度图，其对每个聚类使用相同的端点和间隔。

  实心红色显示表示聚类分布，而颜色较淡的显示则表示总体数据。
- **相对分布**。显示特征名称/标签和单元格中的相对分布。总体而言，显示类似于绝对分布的显示，不同之处在于所显示的是相对分布。

  实心红色显示表示聚类分布，而颜色较淡的显示则表示总体数据。
- **基本视图**。如果聚类很多，不滚动很难看到所有详细信息。要减少滚动量，选择此视图将显示更改为更紧凑的表格。

聚类预测变量重要性视图

“预测变量重要性”视图显示评估模型时每个字段的相对重要性。

聚类大小视图

“聚类大小”视图显示包含每个聚类的饼图。每个聚类的百分比大小显示在每个分区上；鼠标悬停在每个分区上显示该分区中的计数。

图表下方的表格列出以下大小信息:

- 最小聚类的大小（总体计数和百分比）。
- 最大聚类的大小（总体计数和百分比）。
- 最大聚类与最小聚类的大小比率。

单元格分布视图

“单元格分布”视图显示您在“聚类”主面板的表格中选择的任意特征单元格数据分布的展开的详图。

聚类比较视图

“聚类比较”视图由网格式布局构成，行中为特征，列中为选定聚类。此视图帮助您更好地理解组成聚类的因素；同时使您能够看到各聚类间的差异，不但与总体数据比较，而且还在彼此之间比较。

选择要显示的聚类，单击“聚类”主面板中聚类列的顶部。使用 Ctrl+单击或 Shift+单击选择或取消选择多个聚类进行比较。

注：您可以选择最多五个要显示的聚类。

聚类以选择时的顺序显示，而字段顺序则由**特征排序方式**选项决定。当您选择**聚类内重要性**时，将始终按总体重要性顺序排序字段。

背景图显示每个特征的总体分布：

- 类别特征显示为点图，其中点的大小代表每个聚类最频繁出现的（模态）类别（按特征）。
- 连续特征显示为箱图，其显示整体中位数和四分位距。

叠放在这些背景视图上的是所选聚类的箱图：

- 对于连续特征，方点标记和水平线表示每个聚类的中位数和四分位距。
- 每个聚类由不同颜色表示，显示在视图顶部。

浏览聚类查看器

“聚类查看器”为交互式显示。您可以：

- 选择字段或聚类以查看更多详细信息。
- 比较聚类以选择感兴趣的项目。
- 更改显示。
- 变换轴。

使用工具栏

您可使用工具栏选项控制在左右两侧面板中显示的信息。您可使用工具栏控件更改显示的方向（从上至下、从左至右或从右至左）。另外，您还可以将查看器重置为缺省设置，并打开对话框以在主面板中指定“聚类”视图的内容。

仅当您在主面板中选择**聚类**视图时，**特征排序方式**、**聚类排序方式**、**单元格**和**显示**选项才可用。请参阅 第 96 页的『聚类视图』主题以获取更多信息。

表 2. 工具栏图标.

图标	主题
	请参阅变换聚类和特征
	请参阅特征排序方式
	请参阅聚类排序方式
	请参阅单元格

控制聚类视图显示

要控制主面板的聚类视图中显示的内容，单击**显示**按钮；打开“显示”对话框。

特征。缺省选定。要隐藏所有输入特征，取消选择该复选框。

评估字段。选择要显示的评估字段（不用于创建聚类模型的字段，但被发送至模型查看器以评估聚类）；缺省不显示任何字段。*注:* 评估字段必须是包含多个值的字符串。如果无评估字段可用，那么此复选框不可用。

聚类描述。缺省选定。要隐藏所有聚类描述单元格，取消选择该复选框。

聚类大小。缺省选定。要隐藏所有聚类大小单元格，取消选择该复选框。

最大类别数。指定在类别特征图表中显示的最大类别数量；缺省值是 20。

过滤记录

如果希望了解有关特定聚类或聚类组中个案的详细信息，您可以选择记录的子集以基于所选聚类进一步进行分析。

1. 在“聚类查看器”的聚类视图中选择多个聚类。要选择多个聚类，使用 Ctrl+单击。
2. 从菜单中选择:

   生成 > 过滤记录...
3. 输入一个过滤变量名称。所选聚类中的记录将接收此字段值 1。其他所有记录将接收值 0，这些记录将从后续分析中排除，直至您更改过滤状态。
4. 单击**确定**。

第 25 章 系统聚类分析

此过程尝试根据选定的特征来识别相对均一的个案（变量）组，使用的算法是从单独聚类中的每个个案（或变量）开始对各聚类进行组合，直至剩下一个类别。您可以分析原始变量，也可以从多种标准化的转换中选择。距离或相似性测量由“近似值”过程生成。每一阶段均显示统计，以帮助您选择最佳的解。

示例。是否有若干可识别的电视节目组能够吸引各组内相似的观众？利用系统聚类分析，您可以根据观众特征将电视节目（个案）聚类为均一的组。这可以用于识别市场分类以开展市场营销活动。您还可以将城市（个案）聚类到均一组中，从而选择可比城市来检验各种市场营销策略。

统计。单个解或一定范围的解的合并进程表、距离（或相似性）矩阵或聚类成员。图：谱系图和冰柱图。

系统聚类分析数据注意事项

数据。变量可以是定量数据、二元数据或计数数据。变量定标是一个重要问题 – 定标之间的差异可能会影响您的聚类解。如果变量在定标上有很大差异（例如：一个变量以美元为单位度量，而另一个以年数为单位度量），那么应考虑对它们进行标准化（这可以通过“系统聚类分析”过程来自动完成）。

个案顺序。如果相同的距离或相似性存在于输入数据中或产生于连接过程中更新的聚类之间，那么作为结果产生的聚类解会取决于文件中个案的顺序。您可能想要通过以不同随机顺序排序的案例来得到多个不同的解，以验证给定解的稳定性。

假设。所用的距离或相似性测量应适合所分析的数据（请参阅“近似值”过程以获取关于选择距离和相似性测量的更多信息）。并且，应在分析中包含所有相关变量。遗漏有影响的变量会产生错误的解。因为系统聚类分析是一种探测性的方法，其结果应被视为试探性的，直至用独立样本加以确认。

获取系统聚类分析

1. 从菜单中选择:

   分析 > 分类 > 系统聚类...

2. 如果您在对个案聚类，请至少选择一个数值变量。如果您在对变量聚类，请至少选择三个数值变量。

另外，您还可以选择标识变量来标记个案。

系统聚类分析方法

聚类方法。可用的选项有组间联接、组内联接、最近邻元素、最远邻元素、质心聚类法、中位数聚类法和 Ward 法。

度量。允许您指定聚类中使用的距离或相似性测量。选择数据类型以及合适的距离或相似性测量:

- **区间**。可用的选项有 Euclidean 距离、平方 Euclidean 距离、余弦、Pearson 相关性、Chebychev、块、Minkowski 及定制。
- **计数**。可用的选项有卡方测量和 phi 平方测量。
- **二分类**。可用的选项有 Euclidean 距离、平方 Euclidean 距离、刻度差分、模式差分、方差、离差、形状、简单匹配、Phi 4 点相关性、lambda、Anderberg 的 *D*、骰子、Hamann、Jaccard、Kulczynski 1、Kulczynski 2、Lance 和 Williams、Ochiai、Rogers 和 Tanimoto、Russel 和 Rao、Sokal 和 Sneath 1、Sokal 和 Sneath 2、Sokal 和 Sneath 3、Sokal 和 Sneath 4、Sokal 和 Sneath 5、Yule 的 *Y* 以及 Yule 的 *Q*。

转换值。允许您在计算近似值之前为个案或值进行数据值标准化（对二分类数据不可用）。可用的标准化方法包括 *z* 得分、范围从 -1 到 1、范围从 0 到 1、最大量级为 1、平均值为 1 以及标准差为 1。

转换测量。允许您转换距离测量所生成的值。在计算了距离测量之后应用这些转换。可用的备用项包括绝对值、更改符号以及重定比例到 0-1 范围。

系统聚类分析统计

合并进程表。显示在每个阶段合并的个案或聚类、所合并的个案或聚类之间的距离以及个案（或变量）与聚类相联结时所在的最后一个聚类级别。

近似值矩阵。给出各项之间的距离或相似性。

聚类成员。显示在合并聚类的一个或多个阶段中，每个个案被分配所属的聚类。可用的选项有单个解和一定范围的解。

系统聚类分析: 图

谱系图。显示*谱系图*。谱系图可用于评估所形成的聚类的凝聚性，并且可以提供关于要保留的适当聚类数目的信息。

冰柱。显示*冰柱图*，包括所有聚类或指定范围内的聚类。冰柱图显示关于在分析的每次迭代时如何将个案合并到聚类的信息。“方向”允许您选择垂直或水平图。

系统聚类分析: 保存新变量

聚类成员。允许您为单个解或一定范围的解保存聚类成员。然后可以在随后的分析中使用所保存的变量来探索各组之间的其他差别。

CLUSTER 命令语法的其他功能

系统聚类过程使用 CLUSTER 命令语法。使用命令语法语言还可以:

- 在单个分析中使用多个聚类方法。
- 读取并分析近似值矩阵。
- 将近似值矩阵写入磁盘以供以后分析。
- 为定制（幂）距离测量中的幂和根指定任何值。
- 指定已保存的变量的名称。

请参阅*命令语法参考*以获取完整的语法信息。

第 26 章 K 平均值聚类分析

此过程使用可以处理大量个案的算法，根据选定的特征尝试对相对均一的个案组进行标识。不过，该算法要求您指定聚类的个数。如果知道，您可以指定初始聚类中心。您可以选择对个案分类的两种方法之一，要么迭代地更新聚类中心，要么只进行分类。可以保存聚类成员、距离信息和最终聚类中心。还可以选择指定一个变量，使用该变量的值来标记个案输出。您还可以请求分析方差 *F* 统计。尽管这些统计是机会性的（此过程尝试形成不同的组），但统计的相对大小可提供有关各变量对组分离情况的贡献的信息。

示例。哪些可识别的电视节目组能够吸引每个组内的相似观众？通过 *K* 平均值聚类分析，您可以根据观看者的特征将电视节目（个案）聚类为 *K* 均一组。此过程可用于识别市场分类以开展市场营销活动。您还可以将城市（个案）聚类到均一组中，从而选择可比城市来检验各种市场营销策略。

统计。完整解：初始聚类中心和 ANOVA 表。每个个案：聚类信息以及与聚类中心的距离。

K 平均值聚类分析数据注意事项

数据。变量应在区间或定比级别上是定量的。如果您的变量是二分类变量或计数变量，那么使用“系统聚类分析”过程。

个案和初始聚类中心顺序。用于选择初始聚类中心的缺省算法对个案顺序不是保持不变的。“迭代”对话框中的**使用运行平均值**选项使结果解与个案顺序潜在相关，而不管初始聚类中心是如何选择的。如果您使用这些方法种的任一种，那么可能要使用以不同的随机顺序排序的个案获取多个不同的解，以验证给出解的稳定性。指定初始聚类中心且不使用**使用运行平均值**选项将避免与个案顺序相关的问题。然而，如果从个案到聚类中心有固定距离，那么初始聚类中心的排序方式可能会影响解。要获得给定解的稳定性，可以将分析的结果与初始中心值的不同排列相比较。

假设。使用简单欧式距离计算距离。如果想要使用其他距离或相似性测量，请使用“系统聚类分析”过程。变量定标是一个重要的注意事项。如果以不同的刻度测量变量（例如一个变量以美元为单位而另一个以年为单位），那么结果可能令人误解。在此类情况下，应考虑在执行 *K* 平均值聚类分析之前对变量进行标准化（此任务可在“描述”过程中完成）。此过程假设您已选择合适数目的聚类，且已包含所有相关变量。如果您选择的聚类数量不合适或者遗漏了重要的变量，那么结果可能令人误解。

获取 K 平均值聚类分析

1. 从菜单中选择:

   分析 > 分类 > K-平均值聚类...
2. 选择要在聚类分析中使用的变量。
3. 指定聚类数目。（聚类数目必须至少为 2，且不能大于数据文件中的个案数。）
4. 选择**迭代与分类**或者**仅分类**。
5. 或者，选择标识变量标注个案。

K 平均值聚类分析有效性

K 平均值聚类分析命令是非常有效的，主要因为它不像许多聚类算法（包括系统聚类命令使用的算法）那样计算所有个案对之间的距离。

© Copyright IBM Corporation 1989, 2014

为获得最佳有效性，可取一个个案样本并选择**迭代和分类**方法确定聚类中心。选择**最终聚类中心另存为**。然后恢复整个数据文件并选择**仅分类**作为方法，并选择**读取初始聚类中心来源**以使用该样本估计的中心对整个文件分类。您可以写入和读取文件或数据集。可以在同一会话中继续使用数据集，但不会将其另存为文件，除非在会话结束之前明确将其保存为文件。数据集名称必须符合变量命名规则。请参阅主题以获取更多信息。

K 平均值聚类分析: 迭代

注: 只有在您从"K 平均值聚类分析"对话框中选择了**迭代和分类**方法的情况下，这些选项才可用。

最大迭代次数。限制 *K* 平均值算法中的迭代次数。即使尚未满足收敛性准则，达到迭代次数之后迭代也会停止。此数字必须在 1 到 999 之间。

要再次使用版本 5.0 以前的 Quick Cluster 命令使用的算法，应将**最大迭代次数**设置为 1。

收敛性标准。确定迭代何时停止。它表示初始聚类中心之间的最小距离的比例，因此必须大于 0 且小于等于 1。例如，如果准则等于 0.02，那么当完整的迭代无法将任何聚类中心移动任意初始聚类中心之间最小距离的 2% 时，迭代停止。

使用运行平均值。允许您请求在分配了每个个案之后更新聚类中心。如果不选择此选项，那么会在分配了所有个案之后计算新的聚类中心。

K 平均值聚类分析: 保存

可将关于解的信息保存为新变量，以便在后续分析中使用:

聚类成员。创建指示每个个案最终聚类成员的新变量。新变量的值范围是从 1 到聚类数。

与聚类中心的距离。创建指示每个个案与其分类中心之间的欧式距离的新变量。

K 平均值聚类分析: 选项

统计。您可以选择下列统计: 初始聚类中心、ANOVA 表以及每个个案的聚类信息。

- *初始聚类中心*。每个聚类的变量平均值的第一个估计值。缺省情况下，从数据中选择与聚类数相等的分布良好的多个个案。初始聚类中心用于第一轮分类，然后再更新。
- *ANOVA 表 (ANOVA table)*. 显示方差分析表，该表包含每个聚类变量的一元 F 检验。F 检验只是描述性的，不应解释生成的概率。如果所有个案均分配到单独一个聚类，那么 ANOVA 表不显示。
- *每个个案的聚类信息*。显示每个个案的最终聚类分配，以及该个案和用来对个案分类的聚类中心之间的 Euclidean 距离。还显示最终聚类中心之间的欧氏距离。

缺失值。可用的选项为**按列表排除个案**或**按对排除个案**。

- **按列表排除个案**。从分析中排除含任意聚类变量缺失值的个案。
- **按对排除个案**。根据从所有具有非缺失值的变量计算得到的距离将个案分配到聚类。

QUICK CLUSTER 命令的附加功能

"K 平均值聚类"过程使用 QUICK CLUSTER 命令语法。使用命令语法语言还可以:

- 接受前 *k* 个个案作为初始聚类中心，这样可避免通常用于估计初始聚类中心的数据传递。
- 作为命令语法的一部分直接指定初始聚类中心。
- 指定已保存的变量的名称。

请参阅*命令语法参考*以获取完整的语法信息。

第 27 章 非参数检验

非参数检验对数据的基础分布做出最小假设。这些对话框中的可用检验可基于数据组织方式分组为三个较大的类别。

- 单样本检验分析单个字段。
- 相关样本检验对同一组个案的两个或更多字段进行比较。
- 独立样本检验可以分析单个字段，该字段按另一字段的类别进行分组。

单样本非参数检验

单样本非参数检验使用一个或多个非参数检验识别单个字段中的差别。非参数检验不假定您的数据呈正态分布。

您的目标是什么？ 目标允许您快速指定常用的不同检验设置。

- **自动比较观察数据和假设数据**。该目标对仅具有两个类别的分类字段应用二项式检验，对所有其他分类字段应用卡方检验，对连续字段应用 Kolmogorov-Smirnov 检验。
- **检验随机序列**。该目标使用游程检验来检验观察到的随机数据值序列。
- **自定义分析**。当您希望手动修改“设置”选项卡上的检验设置时，选中此选项。注意，如果您随后在“设置”选项卡上更改了与当前选定目标不一致的选项，那么会自动选择该设置。

获取单样本非参数检验

从菜单中选择:

分析 > 非参数检验 > 单样本...

1. 单击**运行**。

根据需要，您可以:

- 在“目标”选项卡上指定目标。
- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。

“字段”选项卡

“字段”选项卡指定应检验哪些字段。

使用预定义角色。此选项使用现有的字段信息。所有预定义角色为“输入”、“目标”或“两者”的字段将用作检验字段。需要至少一个检验字段。

使用自定义字段分配。此选项允许您覆盖字段角色。选定该选项后，指定如下字段:

- **检验字段**。选择一个或更多字段。

“设置”选项卡

“设置”选项卡包含若干组不同的设置，您可以对其进行修改以微调算法如何处理数据。如果您对与当前选定目标不一致的缺省设置进行了更改，那么“目标”选项卡会自动更新为选择**自定义分析**选项。

© Copyright IBM Corporation 1989, 2014

选择检验

这些设置指定要在“字段”选项卡上所指定的字段上执行的检验。

根据数据自动选择检验。该设置对仅具有两个有效（非缺失）类别的分类字段应用二项式检验，对所有其他分类字段应用卡方检验，对连续字段应用 Kolmogorov-Smirnov 检验。

自定义检验。这些设置允许您选择要执行的特定检验。

- **比较观察二分类可能性和假设二分类可能性（二项式检验）**。二项式检验可以应用到所有字段。这将生成一个单样本检验，可以检验标记字段（只有两个类别的分类字段）的观察分布是否与指定的二项式分布期望相同。此外，您还可以请求置信区间。有关检验设置的详细信息，请参阅『二项式检验选项』。
- **比较观察可能性和假设可能性（卡方检验）**。卡方检验可以应用到名义和有序字段。这将生成一个单样本检验，它可以根据字段类别的观察和期望频率间的差异来计算卡方统计。有关检验设置的详细信息，请参阅『卡方检验选项』。
- **检验观察分布和假设分布（Kolmogorov-Smirnov 检验）**。Kolmogorov-Smirnov 检验适用于连续字段和有序字段。这将生成一个单样本检验，即字段的样本累积分布函数是否为齐次的均匀分布、正态分布、泊松分布或指数分布。有关检验设置的详细信息，请参阅第 109 页的『Kolmogorov-Smirnov 选项』。
- **比较中位数和假设中位数（Wilcoxon 带符号等级检验）**。Wilcoxon 带符号等级检验适用于连续字段和有序字段。这将生成一个字段中值的单样本检验。指定一个数字作为假设中位数。
- **检验随机序列（游程检验）**。游程检验可以应用到所有字段。这将生成一个单样本检验，即对分字段的值序列是否为随机序列。有关检验设置的详细信息，请参阅第 109 页的『游程检验选项』。

二项式检验选项: 二项式检验适用于标记字段（只有两个类别的分类字段），但可通过使用定义“成功”的规则应用到所有字段。

假设比例。这指定定义为“成功”的记录的期望比例，即 *p*。请指定大于 0 且小于 1 的值。缺省值为 0.5。

置信区间。可以使用以下方法计算二分类数据的置信区间:

- **Clopper-Pearson（精确）**。基于累积二项式分布的精确区间。
- **Jeffreys**。基于 *p* 的后验分布且应用 Jeffreys 先验的 Bayesian 区间。
- **似然比**。基于 *p* 的似然函数的区间。

定义分类字段的成功。这可以指定如何为分类字段定义对照假设比例检验数据值的“成功”。

- **使用在数据中找到的第一个类别**将使用在样本中找到的第一个定义“成功”的值执行二项式检验。此选项仅适用于只有两个值的名义或有序字段；如果使用了此选项，那么在“字段”选项卡中指定的所有其他分类字段都不会检验。这是缺省值。
- **指定成功值**将使用指定以定义“成功”的值列表来执行二项式检验。可以指定字符串或数值列表。列表中的值不需要在样本中出现。

定义连续字段的成功值。这可以指定如何为连续字段定义对照检验值检验数据值的“成功”。成功被定义为等于或小于分割点的值。

- **样本中点**在最小值和最大值的平均值上设置分割点。
- **自定义分割点**允许您为分割点指定一个值。

卡方检验选项: **所有类别具有相等的概率**。这将在样本中的所有类别间生成均等的频率。这是缺省值。

自定义期望可能性。这允许您为指定的类别列表指定不相等的频率。可以指定字符串或数值列表。列表中的值不需要在样本中出现。在**类别**列中，指定类别值。在**相对频率**列中，为每个类别指定一个大于 0 的值。自定义

的频率被视为比率，例如，指定频率 1、2 和 3 等同于指定频率 10、20 和 30，两者均指定了期望 1/6 的记录属于第一个类别，1/3 的记录属于第二个类别，1/2 的记录属于第三个类别。在指定自定义期望可能性时，自定义类别值必须包括数据中的所有字段值；否则将不对该字段执行检验。

Kolmogorov-Smirnov 选项: 此对话框指定应当检验哪些分布和假设分布的参数。

正态。 **使用样本数据**使用观察到的平均值和标准差；**自定义**允许您指定值。

均匀。 **使用样本数据**使用观察到的最小值和最大值；**自定义**允许您指定值。

指数。 **样本平均值**使用观察到的平均值；**自定义**允许您指定值。

泊松。 **样本平均值**使用观察到的平均值；**自定义**允许您指定值。

游程检验选项: 游程检验适用于标记字段（只有两个类别的分类字段），但可通过使用定义组的规则应用到所有字段。

定义分类字段的组。可用选项有:

- **样本中仅有 2 个类别**使用在定义组的样本中找到的值来执行游程检验。此选项仅适用于只有两个值的名义或有序字段；如果使用了此选项，那么在“字段”选项卡中指定的所有其他分类字段都不会检验。
- **将数据重新编码为 2 个类别**使用指定以定义某个组的值列表来执行游程检验。样本中的所有其他值定义其他组。列表中的值不需要在样本中出现，但每个组中必须至少有一条记录。

定义连续字段的分割点。这可以指定如何为连续字段定义组。第一组定义为等于或小于分割点的值。

- **样本中位数**在样本中位数处设置分割点。
- **样本平均值**在样本平均值处设置分割点。
- **自定义**允许您为分割点指定一个值。

检验选项

显著性水平。这可以指定所有检验的显著性水平 (alpha)。请指定介于 0 和 1 之间的数值。缺省值为 0.05。

置信区间 (%)。这可以指定所有生成的置信区间的置信度。请指定介于 0 与 100 之间的数字值。缺省值为 95。

已排除的个案。这可以指定如何确定检验的个案基础。

- **按列表排除个案**表示从所有分析中排除在“字段”选项卡上指定的任何字段中具有缺失值的记录。
- **按检验排除个案检验**表示从特定检验中排除在此检验所使用字段中具有缺失值的记录。如果在分析中指定了多个检验，将分别独立计算每个检验。

用户缺失值

分类字段的用户缺失值。要在分析中包含记录，分类字段必须具有有效值。通过这些控制可以决定是否将用户缺失值在分类字段中视为有效值。系统缺失值和连续字段缺失值总是被视为无效。

NPTESTS 命令的附加功能

使用命令语法语言还可以:

- 指定在过程的一次运行中的单样本、独立样本和相关样本检验。

请参阅*命令语法参考*以获取完整的语法信息。

独立样本非参数检验

独立样本非参数检验使用一个或多个非参数检验识别两个或更多组间的差别。非参数检验不假定您的数据呈正态分布。

您的目标是什么？ 目标允许您快速指定常用的不同检验设置。

- **自动比较不同组间的分布**。该目标将对具有两个组的数据应用 Mann-Whitney U 检验，或对具有 *k* 个组的数据应用 Kruskal-Wallis 单因素 ANOVA 检验。
- **比较不同组间的中位数**。该目标使用中位数检验来比较在不同组间观察到的中位数。
- **自定义分析**。当您希望手动修改“设置”选项卡上的检验设置时，选中此选项。注意，如果您随后在“设置”选项卡上更改了与当前选定目标不一致的选项，那么会自动选择该设置。

获取独立样本非参数检验

从菜单中选择:

分析 > 非参数检验 > 独立样本...

1. 单击**运行**。

根据需要，您可以:

- 在“目标”选项卡上指定目标。
- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。

“字段”选项卡

“字段”选项卡指定应检验哪些字段和用于定义组的字段。

使用预定义角色。此选项使用现有的字段信息。所有预定义角色为“目标”或“两者”的连续字段和有序字段都将用作检验字段。如果有一个具有预定义角色“输入”的分类字段，它将用作分组字段。否则，缺省不使用分组字段，您必须使用自定义字段分配。需要至少一个检验字段和分组字段。

使用自定义字段分配。此选项允许您覆盖字段角色。选定该选项后，指定如下字段:

- **检验字段**。选择一个或多个连续字段或有序字段。
- **组**。选择一个分类字段。

“设置”选项卡

“设置”选项卡包含若干组不同的设置，您可以对其进行修改以微调算法如何处理数据。如果您对与当前选定目标不一致的缺省设置进行了更改，那么“目标”选项卡会自动更新为选择**自定义分析**选项。

选择检验

这些设置指定要在“字段”选项卡上所指定的字段上执行的检验。

根据数据自动选择检验。该设置将对具有两个组的数据应用 Mann-Whitney U 检验，或对具有 *k* 个组的数据应用 Kruskal-Wallis 单因素 ANOVA 检验。

自定义检验。这些设置允许您选择要执行的特定检验。

- **比较不同组间的分布**。这些将生成独立样本检验，即样本是否来自同一总体。

Mann-Whitney U（二样本） 使用每个个案的等级来检验组是否抽取自同一总体。分组字段中按升序排列的第一个值定义第一个组，第二个值定义第二个组。如果分组字段有两个以上的值，那么不生成此检验。

Kolmogorov-Smirnov（二样本）对两个分布间中位数、离散、偏度等的任何差异很敏感。如果分组字段有两个以上的值，那么不生成此检验。

检验随机序列（二样本 Wald-Wolfowitz）生成一个以组成员关系为准则的游程检验。如果分组字段有两个以上的值，那么不生成此检验。

Kruskal-Wallis 单因素 ANOVA（k 个样本）是 Mann-Whitney U 检验的扩展，它也是单向方差检验分析的非参数模拟。您可以根据需要请求对 *k* 样本的多重比较，即**所有成对**多重比较或**逐步降低**比较。

有序选项检验（k 样本 Jonckheere-Terpstra）可作为比 Kruskal-Wallis 功能更强大的选项，但前提是 *k* 样本需具有自然顺序。例如，*k* 个总体可能代表 *k* 个上升的温度。“不同的温度产生相同的响应分布”这一假设是针对“温度升高，那么响应的量级增加”这一选择进行检验的。此处备选假设已排序，因此，Jonckheere-Terpstra 是最适用的检验。**从最小到最大**指定其他假设，第一组的位置参数小于或等于第二组，而第二组又小于或等于第三组，以此类推。**从最大到最小**指定其他假设，第一组的位置参数大于或等于第二组，而第二组又大于或等于第三组，以此类推。对于这两个选项，其他假设还假定位置不相等。您可以根据需要请求对 *k* 样本的多重比较，即**所有成对**多重比较或**逐步降低**比较。

- **比较不同组间的范围**。这可以生成一个独立样本检验，即样本是否具有相同范围。**Moses 极端反应（二样本）**检验控制组与比较组。分组字段中按升序排列的第一个值定义控制组，第二个值定义比较组。如果分组字段有两个以上的值，那么不生成此检验。
- **比较不同组间的中位数**。这可以生成一个独立样本检验，即样本是否具有相同中位数。**中位数检验（k 个样本）**可以使用汇聚样本中位数（从数据集所有记录中计算）或自定义值作为假设中位数。您可以根据需要请求对 *k* 样本的多重比较，即**所有成对**多重比较或**逐步降低**比较。
- **估计不同组间的置信区间**。 **Hodges-Lehman 估计（二样本）**可以为两个组的中位数差异生成一个独立样本估计和置信区间。如果分组字段有两个以上的值，那么不生成此检验。

检验选项

显著性水平。这可以指定所有检验的显著性水平 (alpha)。请指定介于 0 和 1 之间的数值。缺省值为 0.05。

置信区间 (%)。这可以指定所有生成的置信区间的置信度。请指定介于 0 与 100 之间的数字值。缺省值为 95。

已排除的个案。这可以指定如何确定检验的个案基础。**按列表排除个案**表示从所有分析中排除在任何子命令上指定的任何字段中具有缺失值的记录。**按检验排除个案检验**表示从特定检验中排除在此检验所使用字段中具有缺失值的记录。如果在分析中指定了多个检验，将分别独立计算每个检验。

用户缺失值

分类字段的用户缺失值。要在分析中包含记录，分类字段必须具有有效值。通过这些控制可以决定是否将用户缺失值在分类字段中视为有效值。系统缺失值和连续字段缺失值总是被视为无效。

NPTESTS 命令的附加功能

使用命令语法语言还可以:

- 指定在过程的一次运行中的单样本、独立样本和相关样本检验。

请参阅*命令语法参考*以获取完整的语法信息。

相关样本非参数检验

使用一个或多个非参数检验识别两个或多个相关字段之间的差别。非参数检验不假定您的数据呈正态分布。

数据注意事项。每个记录对应于有两个或更多相关测量值存储在数据集中单独字段中的给定受试人。例如，如果每个受试人的体重以定期间隔测量并存储在如*节食前体重*、*中间体重*和*节食后体重*这样的字段中，那么可使用样本相关非参数检验分析节食计划的有效性研究。这些字段为“相关”。

您的目标是什么？ 目标允许您快速指定常用的不同检验设置。

- **自动比较观察数据和假设数据**。当指定 2 个字段时，该目标对分类数据应用 McNemar 检验；当指定超过 2 个字段时，那么对分类数据应用 Cochran 的 Q 检验；当指定 2 个字段时，对连续数据应用 Wilcoxon 匹配对符号等级检验；当指定超过 2 个字段时，对连续数据应用 Friedman 的按等级二因素 ANOVA 检验。
- **自定义分析**。当您希望手动修改“设置”选项卡上的检验设置时，选中此选项。注意，如果您随后在“设置”选项卡上更改了与当前选定目标不一致的选项，那么会自动选择该设置。

当指定了不同测量级别的字段时，它们首先由测量级别隔开，然后将相应检验应用到每个组。例如，如果您选择**自动比较观测数据和假设数据**作为您的目标，并指定 3 个连续字段和 2 个名义字段，那么会将 Friedman 检验应用到连续字段并将 McNemar 检验应用到名义字段。

获取相关样本非参数检验

从菜单中选择:

分析 > 非参数检验 > 相关样本...

1. 单击**运行**。

根据需要，您可以:

- 在“目标”选项卡上指定目标。
- 在“字段”选项卡上指定字段分配。
- 在“设置”选项卡上指定专家设置。

“字段”选项卡

“字段”选项卡指定应检验哪些字段。

使用预定义角色。此选项使用现有的字段信息。所有预定义角色为“目标”或“两者”的字段将用作检验字段。需要至少两个检验字段。

使用自定义字段分配。此选项允许您覆盖字段角色。选定该选项后，指定如下字段:

- **检验字段**。选择两个或更多字段。每个字段对应一个单独的相关样本。

“设置”选项卡

“设置”选项卡包含若干组不同的设置，您可以对其进行修改以微调过程如何处理数据。如果您对与其他目标不一致的缺省设置进行了更改，那么“目标”选项卡会自动更新为选择**自定义分析**选项。

选择检验

这些设置指定要在“字段”选项卡上所指定的字段上执行的检验。

根据数据自动选择检验。当指定 2 个字段时，此设置对分类数据应用 McNemar 检验；当指定超过 2 个字段时，那么对分类数据应用 Cochran 的 Q 检验；当指定 2 个字段时，对连续数据应用 Wilcoxon 匹配对符号等级检验；当指定超过 2 个字段时，对连续数据应用 Friedman 的按等级二因素 ANOVA 检验。

自定义检验。这些设置允许您选择要执行的特定检验。

- **检验二分类数据中的更改。 McNemar 检验（二样本）**可以应用于分类字段。这将生成一个相关样本检验，即两个标记字段（只有两个值的分类字段）间的值组合可能性是否相同。如果在“字段”选项卡上指定两个以上的字段，将不执行此检验。有关检验设置的详细信息，请参阅『McNemar 检验: 定义成功』。**Cochran 的 Q（k 个样本）**可以应用到分类字段。这将生成一个相关样本检验，即 *k* 个标记字段（只有两个值的分类字段）间的值组合可能性是否相同。您可以根据需要请求对 *k* 样本的多重比较，即**所有成对**多重比较或**逐步降低**比较。有关检验设置的详细信息，请参阅『Cochran Q: 定义成功』。
- **检验多项数据中的更改。 边际同质性检验（二样本）**生成一个相关样本检验，即，检验两个配对有序字段之间的值组合的可能性是否相同。边际同质性检验通常在重复测量情况下使用。此检验是 McNemar 检验从二值响应到多项式响应的扩展。如果在“字段”选项卡上指定两个以上的字段，将不执行此检验。
- **比较中位数差和假设中位数差**。这些检验各生成一个相关样本检验，即，检验两个字段间的中位数差是否不等于 0。此检验适用于连续字段和有序字段。如果在“字段”选项卡上指定了两个以上的字段，那么将不会执行这些检验。
- **估计置信区间**。这将为两个配对字段间的中位数差生成一个相关样本估计和置信区间。此检验适用于连续字段和有序字段。如果在“字段”选项卡上指定两个以上的字段，将不执行此检验。
- **量化关联。 Kendall 协同系数（K 个样本）**将生成对裁判员或评分者间一致性的测量，每条记录为单个裁判员对多个项目（字段）的评价。您可以根据需要请求对 *k* 样本的多重比较，即**所有成对**多重比较或**逐步降低**比较。
- **比较分布。 Friedman 按等级二因素 ANOVA（k 个样本）**将生成一个相关样本检验，即，检验 *k* 个相关样本是否从同一群体中抽取。您可以根据需要请求对 *k* 样本的多重比较，即**所有成对**多重比较或**逐步降低**比较。

McNemar 检验: 定义成功: McNemar 检验适用于标记字段（只有两个类别的分类字段），但可通过使用定义“成功”的规则应用到所有分类字段。

定义分类字段的成功。这可以指定如何为分类字段定义“成功”。

- **使用在数据中找到的第一个类别**将使用在样本中找到的第一个定义“成功”的值执行检验。此选项仅适用于只有两个值的名义或有序字段；如果使用了此选项，那么在“字段”选项卡中指定的所有其他分类字段都不会检验。这是缺省值。
- **指定成功值**将使用指定以定义“成功”的值列表来执行检验。可以指定字符串或数值列表。列表中的值不需要在样本中出现。

Cochran Q: 定义成功: Cochran 的 Q 检验适用于标记字段（只有两个类别的分类字段），但可通过使用定义“成功”的规则应用到所有分类字段。

定义分类字段的成功。这可以指定如何为分类字段定义“成功”。

- **使用在数据中找到的第一个类别**将使用在样本中找到的第一个定义“成功”的值执行检验。此选项仅适用于只有两个值的名义或有序字段；如果使用了此选项，那么在“字段”选项卡中指定的所有其他分类字段都不会检验。这是缺省值。
- **指定成功值**将使用指定以定义“成功”的值列表来执行检验。可以指定字符串或数值列表。列表中的值不需要在样本中出现。

检验选项

显著性水平。这可以指定所有检验的显著性水平 (alpha)。请指定介于 0 和 1 之间的数值。缺省值为 0.05。

置信区间 (%)。这可以指定所有生成的置信区间的置信度。请指定介于 0 与 100 之间的数字值。缺省值为 95。

已排除的个案。这可以指定如何确定检验的个案基础。

- **按列表排除个案**表示从所有分析中排除在任何子命令上指定的任何字段中具有缺失值的记录。
- **按检验排除个案检验**表示从特定检验中排除在此检验所使用字段中具有缺失值的记录。如果在分析中指定了多个检验，将分别独立计算每个检验。

用户缺失值

分类字段的用户缺失值。要在分析中包含记录，分类字段必须具有有效值。通过这些控制可以决定是否将用户缺失值在分类字段中视为有效值。系统缺失值和连续字段缺失值总是被视为无效。

NPTESTS 命令的附加功能

使用命令语法语言还可以:

- 指定在过程的一次运行中的单样本、独立样本和相关样本检验。

请参阅*命令语法参考*以获取完整的语法信息。

模型视图

模型视图

此过程在查看器中创建“模型查看器”对象。激活（双击）该对象，可获得模型的交互式视图。此模型视图具有一个双面板窗口，主视图位于左侧，链接或辅助视图位于右侧。

有两个主视图:

- 假设摘要。这是缺省视图。请参阅『假设摘要』主题以获取更多信息。
- 置信区间摘要。请参阅 第 115 页的『置信区间摘要』主题以获取更多信息。

有七个链接/辅助视图:

- 单样本检验。如果请求单样本检验，这是缺省视图。请参阅第 115 页的『单样本检验』主题以获取更多信息。
- 相关样本检验。如果请求了相关样本检验，同时未请求单样品检验，这是缺省视图。请参阅第 116 页的『相关样本检验』主题以获取更多信息。
- 独立样本检验。如果未请求相关样本检验或单样本检验，这是缺省视图。请参阅第 117 页的『独立样本检验』主题以获取更多信息。
- 分类字段信息。请参阅第 118 页的『分类字段信息』主题以获取更多信息。
- 连续字段信息。请参阅第 118 页的『连续字段信息』主题以获取更多信息。
- 成对比较。请参阅第 118 页的『成对比较』主题以获取更多信息。
- 齐次子集。请参阅 第 118 页的『均一子集』主题以获取更多信息。

假设摘要

“模型摘要”视图是非参数检验的快照摘要。它强调原假设和决策，引起对显著 *p* 值的注意。

- 每一行对应一个单独的检验。单击行在链接视图中显示有关该检验的附加信息。

- 单击任何列标题将按该列的值排序行。
- 使用**重置**按钮可以将“模型查看器”恢复到其原始状态。
- 使用**字段过滤器**下拉列表可以只显示涉及所选字段的检验。

置信区间摘要

“置信区间摘要”显示非参数检验生成的任何置信区间。

- 每一行对应一个单独的置信区间。
- 单击任何列标题将按该列的值排序行。

单样本检验

“单样本检验”视图显示与任何请求的单样本非参数检验相关的详细信息。显示的信息取决于所选的检验。

- 使用**检验**下拉列表可以选择给定类型的单样本检验。
- 使用**字段**下拉列表可以选择经**检验**下拉列表中所选检验检验过的字段。

二项式检验

“二项式检验”显示堆积条形图和检验表。

- 堆积条形图显示检验字段“成功”和“失败”类别的观察频率和假设频率，其中“失败”堆积在“成功”的顶部。悬停在条形上将在工具提示中显示类别百分比。条形中的可见区别表明检验字段可能没有假设的二项式分布。
- 该表显示检验的详细信息。

卡方检验

“卡方检验”视图显示复式条形图和检验表。

- 复式条形图显示检验字段每个类别的观察频率和假设频率。悬停在条形上将在工具提示中显示观察频率和假设频率及其差别（残差）。观察和假设条形中的可见区别表明检验字段可能没有假设的分布。
- 该表显示检验的详细信息。

Wilcoxon 带符号等级

“Wilcoxon 带符号等级检验”视图显示直方图和检验表。

- 直方图包括显示观察和假设中位数的垂直线。
- 该表显示检验的详细信息。

游程检验

“游程检验”视图显示图表和检验表。

- 该图表显示以垂直线标记的观察到的游程数的正态分布。注意，当执行精确检验时，该检验不基于正态分布。
- 该表显示检验的详细信息。

Kolmogorov-Smirnov 检验

“Kolmogorov-Smirnov 检验”视图显示直方图和检验表。

- 该直方图包括假设均匀、正态、泊松或指数分布概率密度函数的重叠。注意，该检验基于累积分布，同时表格中报告的“最极端差分”应相对于累积分布进行解释。

- 该表显示检验的详细信息。

相关样本检验

“单样本检验”视图显示与任何请求的单样本非参数检验相关的详细信息。显示的信息取决于所选的检验。

- 使用**检验**下拉列表可以选择给定类型的单样本检验。
- 使用**字段**下拉列表可以选择经**检验**下拉列表中所选检验检验过的字段。

McNemar 检验

“McNemar 检验”视图显示复式条形图和检验表。

- 复式条形图显示检验字段定义的 2×2 表的非对角线单元格的观察频率和假设频率。
- 该表显示检验的详细信息。

符号检验

“符号检验”视图显示堆积直方图和检验表。

- 堆积直方图通过将差别符号用作堆积字段来显示字段间的差别。
- 该表显示检验的详细信息。

Wilcoxon 带符号等级检验

“Wilcoxon 带符号等级检验”视图显示堆积直方图和检验表。

- 堆积直方图通过将差别符号用作堆积字段来显示字段间的差别。
- 该表显示检验的详细信息。

边际同质性检验

“边际同质性检验”视图显示复式条形图和检验表。

- 复式条形图显示检验字段定义的表格的非对角线单元格的观察频率和假设频率。
- 该表显示检验的详细信息。

Cochran 的 Q 检验

“Cochran 的 Q 检验”视图显示堆积条形图和检验表。

- 堆积条形图显示检验字段“成功”和“失败”类别的观察频率，其中“失败”堆积在“成功”的顶部。悬停在条形上将在工具提示中显示类别百分比。
- 该表显示检验的详细信息。

Friedman 的双向按等级方差分析

“Friedman 的双向按等级方差分析”视图显示面板直方图和检验表。

- 该直方图显示观察到的等级分布，按检验字段生成面板。
- 该表显示检验的详细信息。

Kendall 协同系数

“Kendall 协同系数”视图显示面板直方图和检验表。

- 该直方图显示观察到的等级分布，按检验字段生成面板。
- 该表显示检验的详细信息。

独立样本检验

“独立样本检验”视图显示与任何请求的独立样本非参数检验相关的详细信息。显示的信息取决于所选的检验。

- 使用**检验**下拉列表可以选择给定类型的独立样本检验。
- 使用**字段**下拉列表可以选择检验并分组经**检验**下拉列表中所选检验检验过的字段组合。

Mann-Whitney 检验

“Mann-Whitney 检验”视图显示人口金字塔图和检验表。

- 人口金字塔图按照分组字段类别显示连续直方图，注意每个组中的记录数量和组的等级平均值。
- 该表显示检验的详细信息。

Kolmogorov-Smirnov 检验

“Kolmogorov-Smirnov 检验”视图显示人口金字塔图和检验表。

- 人口金字塔图按照分组字段类别显示连续直方图，注意每个组中的记录数量。单击**累积**按钮可以显示或隐藏观察到的累积分布线。
- 该表显示检验的详细信息。

Wald-Wolfowitz 游程检验

“Wald-Wolfowitz 游程检验”视图显示堆积条形图和检验表。

- 人口金字塔图按照分组字段类别显示连续直方图，注意每个组中的记录数量。
- 该表显示检验的详细信息。

Kruskal-Wallis 检验

“Kruskal-Wallis 检验”视图显示箱图和检验表。

- 为分组字段的每个类别显示单独的箱图。悬停在箱上将在工具提示中显示等级平均值。
- 该表显示检验的详细信息。

Jonckheere-Terpstra 检验

“Jonckheere-Terpstra 检验”视图显示箱图和检验表。

- 为分组字段的每个类别显示单独的箱图。
- 该表显示检验的详细信息。

Moses 极端反应检验

“Moses 极端反应检验”视图显示箱图和检验表。

- 为分组字段的每个类别显示单独的箱图。单击**记录 ID** 按钮可以显示或隐藏点标签。
- 该表显示检验的详细信息。

中位数检验

“中位数检验”视图显示箱图和检验表。

- 为分组字段的每个类别显示单独的箱图。
- 该表显示检验的详细信息。

分类字段信息

“分类字段信息”视图显示在**字段**下拉列表上选择的分类字段的条形图。可用字段列表限于在“假设摘要”视图的当前选定检验中使用的分类字段。

- 悬停在条形上将在工具提示中显示类别百分比。

连续字段信息

“连续字段信息”视图显示在**字段**下拉列表上选择的连续字段的直方图。可用字段列表限于在“假设摘要”视图的当前选定检验中使用的连续字段。

成对比较

“成对比较”视图显示在请求了成对多重比较时，*k* 样本非参数检验生成的距离网络图和比较表。

- 距离网络图是比较表的图形表示，其中网络中节点间的距离对应于样本间的差别。黄线对应于统计上的显著差异；黑线对应于不显著的差异。悬停在网络中的直线上将显示具有该直线连接的节点间的调整差异显著性的工具提示。
- 比较表显示所有成对比较的数值结果。每一行对应一个单独的成对比较。单击列标题将按该列的值排序行。

均一子集

“均一子集”视图在请求逐步降低多重比较时，显示 *K* 样本非参数检验生成的比较表。

- “样本”组中的每一行对应于单独的相关样本（以单独字段数据表示）。统计上没有显著性差别的样本被分组到相同颜色的子集中；每个已标记子集都有单独的列。当所有样本具有统计上的显著性差别时，每个样本有一个单独的子集。当没有样本存在统计上的显著性差别时，只有一个子集。
- 为每个包含超过一个样本的子集计算检验统计、显著性值和调整显著性值。

NPTESTS 命令的附加功能

使用命令语法语言还可以:

- 指定在过程的一次运行中的单样本、独立样本和相关样本检验。

请参阅*命令语法参考*以获取完整的语法信息。

传统对话框

有一些“legacy”对话框也可执行非参数检验。这些对话框支持“精确检验”选项提供的功能。

卡方检验。可将一个变量以表格形式列在不同的类别中，并根据观察的和期望的频率之间的差来计算卡方统计。

二项式检验。将二分变量的每个类别中的观察频率与二项式分布中的期望频率进行比较。

游程检验。检验变量的两个值的出现顺序是否随机。

单样本 Kolmogorov-Smirnov 检验。将观测到的变量累积分布函数与指定的理论分布进行比较，该理论分布可以是正态分布、均匀分布、指数分布或泊松分布。

两个独立样本检验。比较一个变量的两组个案。提供了 Mann-Whitney *U* 检验、双样本 Kolmogorov-Smirnov 检验、Moses 极端反应检验和 Wald-Wolfowitz 游程检验。

两个相关样本检验。比较两个变量的分布。提供了 Wilcoxon 带符号等级检验、符号检验和 McNemar 检验。

多个独立样本检验。比较一个变量的两组或更多组个案。提供了 Kruskal-Wallis 检验、中位数检验和 Jonckheere-Terpstra 检验。

多个关联样本检验。比较两个或更多个变量的分布。提供了 Friedman 检验、Kendall *W* 检验和 Cochran *Q* 检验。

四分位数和平均值、标准差、最小值、最大值和非缺失个案数对以上所有检验均可用。

卡方检验

卡方检验过程可将一个变量以表格形式列在不同的类别中，并计算卡方统计。此拟合优度检验比较每个类别中的观察的和期望的频率，以检验所有类别是否包含相同比例的值，或检验每个类别是否包含用户指定比例的值。

示例。卡方检验可用于确定一袋糖豆是否包含相等比例的蓝色、棕色、绿色、橙色、红色和黄色糖果。也可以检验一袋糖豆是否包含 5% 蓝色、30% 棕色、10% 绿色、20% 橙色、15% 红色和 15% 黄色的糖果。

统计。平均值、标准差、最小值、最大值和四分位数。非缺失和缺失个案的数量和百分比；每个类别观测的和期望的个案的个数；残差以及卡方统计。

卡方检验数据注意事项

数据。使用排序的或未排序的数值分类变量（有序或名义测量级别）。要将字符串变量转换为数值变量，请使用“自动重新编码”过程（在“转换”菜单上提供）。

假设。非参数检验不要求假定基础分布的形状。数据均假定为随机样本。每个类别的期望频率应至少为 1。应有不超过 20% 的类别具有小于 5 的期望频率。

获取卡方检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > 卡方...

2. 选择一个或多个检验变量。每个变量产生一个单独的检验。
3. 或者，单击**选项**以获取描述统计、四分位数和缺失数据的处理控制。

卡方检验的期望范围和期望值

期望范围。缺省情况下，变量的每个不同的值均定义为一个类别。要在特定范围内建立类别，请选择**使用指定范围**，并为下限和上限输入整数值。为包含范围内的每个整数值建立类别，并排除具有界外值的个案。例如，如果指定下限值为 1，上限值为 4，那么对卡方检验仅使用 1 到 4 的整数值。

期望值。缺省情况下，所有类别都具有相等的期望值。类别可以包含用户指定的期望比例。选择**值**，为检验变量的每个类别输入一个大于 0 的值，然后单击**添加**。每次添加值时，该值就会出现在值列表的底部。值的顺序很重要；该顺序与检验变量的类别值的升序相对应。列表中的第一个值与检验变量的最低组值相对应，而列表中的最后一个值与最高值相对应。对值列表的元素进行求和，然后每个值除以此和，以计算出相应类别中所期望的个案比例。例如，值列表 3、4、5、4 就指定了期望比例 3/16、4/16、5/16 和 4/16。

卡方检验选项

统计。可以选择一个或全部两个汇总统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。
- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

缺失值。控制对缺失值的处理。

- **按检验排除个案**。如果指定多个检验，将分别独立计算每个检验中的缺失值。
- **按列表排除个案**。从所有分析中排除任何变量具有缺失值的个案。

NPAR TESTS 命令的附加功能（卡方检验）

使用命令语法语言还可以:

- 为不同的变量指定不同的最小值和最大值或期望频率（使用 CHISQUARE 命令）。
- 根据不同的期望频率检验同一变量，或使用不同的范围（使用 EXPECTED 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

二项式检验

二项式检验过程比较二分变量的两个类别的观察频率与指定概率参数的二项式分布下的期望频率。缺省情况下，两个组的概率参数均为 0.5。要更改这两个概率，您可以为第一个组输入一个检验比例。第二个组的概率将是 1 减去第一个组的指定概率。

示例。当您掷出一枚硬币，正面朝上的概率等于 1/2。根据这一假设，将硬币抛掷 40 次，并记录结果（正面朝上和反面朝上的情况）。从二项式检验中，您可能发现，3/4 的抛掷都是正面朝上，并且观测的显著性水平很小 (0.0027)。这些结果表明，正面朝上的概率不可能等于 1/2；硬币可能是有偏倚的。

统计。平均值、标准差、最小值、最大值、非缺失个案数和四分位数。

二项式检验数据注意事项

数据。检验的变量应为数值二分变量。要将字符串变量转换为数值变量，请使用“自动重新编码”过程（在“转换”菜单上提供）。**二分变量**是只能取两个可能值的变量：*yes* 或 *no*，*true* 或 *false*，0 或 1，等等。在数据集中遇到的第一个值定义第一个组，其他值定义第二个组。如果变量不是二分变量，那么必须指定分割点。分割点将具有小于或等于分割点的值的个案指派到第一个组，并将其余个案指派到第二个组。

假设。非参数检验不要求假定基础分布的形状。数据均假定为随机样本。

获取二项式检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > 二项式...

2. 选择一个或多个数值检验变量。
3. 或者，单击**选项**以获取描述统计、四分位数和缺失数据的处理控制。

二项式检验选项

统计。可以选择一个或全部两个汇总统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。
- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

缺失值。控制对缺失值的处理。

- **按检验排除个案**。如果指定多个检验，将分别独立计算每个检验中的缺失值。
- **按列表排除个案**。从所有分析中排除所检验的任何变量具有缺失值的个案。

NPAR TESTS 命令的附加功能（二项式检验）

使用命令语法语言还可以:

- 变量有两个以上类别时，选择特定的组（并排除其他组）（用 BINOMIAL 子命令）。
- 为不同的变量指定不同的分割点或概率（使用 BINOMIAL 子命令）。
- 对照不同的分割点或概率检验同一变量（使用 EXPECTED 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

游程检验

“游程检验”过程检验某一变量的两个值的出现顺序是否随机。游程是相似的观察值的一个序列。游程太多或太少的样本不是随机样本。

示例。假设对 20 个人进行民意调查，以弄清他们是否将购买某产品。如果所有这 20 人均是相同性别，那么假定的样本随机性将受到严重质疑。可以使用游程检验来确定样本是否是随机抽取的。

统计。平均值、标准差、最小值、最大值、非缺失个案数和四分位数。

游程检验数据注意事项

数据。变量必须是数值。要将字符串变量转换为数值变量，请使用“自动重新编码”过程（在“转换”菜单上提供）。

假设。非参数检验不要求假定基础分布的形状。而是使用来自连续概率分布的样本。

获取游程检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > 游程...
2. 选择一个或多个数值检验变量。
3. 或者，单击**选项**以获取描述统计、四分位数和缺失数据的处理控制。

游程检验: 分割点

分割点。指定一个分割点以对分已选择的变量。可以使用观察到的平均值、中位数或众数作为分割点，也可以使用指定的值作为分割点。值小于分割点的个案分配到一个组，值大于或等于分割点的个案分配到另一个组。为每个选择的分割点执行一次检验。

游程检验选项

统计。可以选择一个或全部两个汇总统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。
- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

缺失值。控制对缺失值的处理。

- **按检验排除个案**。如果指定多个检验，将分别独立计算每个检验中的缺失值。
- **按列表排除个案**。从所有分析中排除任何变量具有缺失值的个案。

NPAR TESTS 命令的附加功能（游程检验）

使用命令语法语言还可以:

- 为不同的变量指定不同的分割点（使用 RUNS 子命令）。

- 对照不同的定制分割点检验相同的变量（使用 RUNS 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

单样本 Kolmogorov-Smirnov 检验

单样本 Kolmogorov-Smirnov 检验过程将变量的观察累积分布函数与指定的理论分布进行比较，该理论分布可以是正态分布、均匀分布、泊松分布或指数分布。Kolmogorov-Smirnov *Z* 由观察累积分布函数和理论累积分布函数之间的最大差分（取绝对值）计算而得。该拟合优度检验检验了观察值是否合理来自指定的分布。

示例。许多参数检验都需要正态分布的变量。单样本 Kolmogorov-Smirnov 检验可用于检验变量（例如 *income*）是否为正态分布。

统计。平均值、标准差、最小值、最大值、非缺失个案数和四分位数。

单样本 Kolmogorov-Smirnov 检验数据注意事项

数据。使用定量变量（定距或者定比测量级别）。

假设。Kolmogorov-Smirnov 检验假设检验分布参数已提前指定。此过程估计样本中的参数。样本平均值和样本标准差是正态分布的参数。样本最小值和最大值定义了均匀分布的极差，样本平均值是泊松分布的参数，样本平均值是指数分布的参数。检测偏离假设分布的检验功能可能严重降低。对于以估计参数对正态分布进行的检验，考虑调整的 K-S Lilliefors 检验（在“探索”过程中）。

获取单样本 Kolmogorov-Smirnov 检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > 单样本 K-S...
2. 选择一个或多个数值检验变量。每个变量产生一个单独的检验。
3. 或者，单击**选项**以获取描述统计、四分位数和缺失数据的处理控制。

单样本 Kolmogorov-Smirnov 检验: 选项

统计。可以选择一个或全部两个汇总统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。
- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

缺失值。控制对缺失值的处理。

- **按检验排除个案**。如果指定多个检验，将分别独立计算每个检验中的缺失值。
- **按列表排除个案**。从所有分析中排除任何变量具有缺失值的个案。

NPAR TESTS 命令的附加功能（单样本 Kolmogorov-Smirnov 检验）

使用命令语法语言还可以指定检验分布的参数（用 K-S 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

两个独立样本检验

“两个独立样本检验”过程根据一个变量来比较两组个案。

示例。新型牙箍已开发出来了，其目的是为了变得更舒适美观，以及提供更快的牙齿矫正过程。要查明戴新牙箍的时间是否必须与旧牙箍一样长，随机选择 10 名儿童戴旧牙箍，再随机选择 10 名儿童戴新牙箍。从 Mann-Whitney *U* 检验中，可能会发现：戴新牙箍的儿童戴牙箍的时间不必与戴旧牙箍的儿童一样长。

统计。平均值、标准差、最小值、最大值、非缺失个案数和四分位数。检验：Mann-Whitney *U*、Moses 极端反应、Kolmogorov-Smirnov *Z* 和 Wald-Wolfowitz 游程。

两个独立样本检验数据注意事项

数据。使用可以排序的数值变量。

假设。使用独立的随机样本。Mann-Whitney *U* 检验检验两个分布的等同性。要使用它来检验两个分布间的位置差别，必须假设二者具有相同的形状。

获取两个独立样本检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > 2 个独立样本...
2. 选择一个或多个数值变量。
3. 选择一个分组变量，并单击**定义组**将文件拆分成两个组或样本。

两个独立样本检验：类型

检验类型。有四种检验可用于检验两个独立样本（组）是否来自同一个总体。

Mann-Whitney U 检验是最常用的两个独立样本检验。它等同于对两个组进行的 Wilcoxon 等级和检验和 Kruskal-Wallis 检验。Mann-Whitney 检验被抽样的两个总体处于等同的位置。对来自两个组的观察值进行组合和等级排序，在同数的情况下分配平均等级。同数的数目相对于观察值总数要小一些。如果两个总体的位置相同，那么在两个样本之间随机混合等级。该检验计算组 1 分数领先于组 2 分数的次数，以及组 2 分数领先于组 1 分数的次数。Mann-Whitney *U* 统计是这两个数字中较小的一个。同时显示 Wilcoxon 等级和 *W* 统计。*W* 是具有较小等级平均值的组的等级之和，除非组具有相同等级平均值，那么它将是在“两个独立样本定义组”对话框中最后命名组的等级之和。

Kolmogorov-Smirnov Z 检验和 **Wald-Wolfowitz 游程检验**是检测分布对于位置和形状的差异所更为通用的检验。Kolmogorov-Smirnov 检验是以两个样本的观察累积分布函数之间的最大绝对差为基础的。当这个差很大时，就将这两个分布视为不同的分布。Wald-Wolfowitz 游程检验对来自两个组的观察值进行组合和等级排序。如果两个样本来自同一总体，那么两个组应随机散布在整个等级中。

Moses 极端反应检验假定实验变量在一个方向影响某些主体，而在相反方向影响其他主体。它检验与控制组相比的极端响应。当与控制组结合时，此检验主要检查控制组的跨度，是对实验组中的极值对该跨度的影响程度的测量。控制组由“两个独立样本：定义组”对话框中的组 1 值定义。来自两个组的观察值都进行了组合和等级排序。控制组的跨度通过控制组的最大值和最小值的等级差加上 1 来计算。因为意外的离群值可能轻易使跨度范围变形，所以将自动从各端修剪 5% 的控制个案。

两个独立样本检验：定义组

要将文件拆分成两个组或样本，请为组 1 输入一个整数值，并为组 2 输入另一个整数值。将具有其他值的个案从分析中排除。

两个独立样本检验：选项

统计。可以选择一个或全部两个汇总统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。

- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

缺失值。控制对缺失值的处理。

- **按检验排除个案**。如果指定多个检验，将分别独立计算每个检验中的缺失值。
- **按列表排除个案**。从所有分析中排除任何变量具有缺失值的个案。

NPAR TESTS 命令的附加功能（两个独立样本检验）

使用命令语句语言还可以指定为 Moses 检验要修剪掉的个案数（用 MOSES 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

两个关联样本检验

两个相关样本检验过程对两个变量的分布进行比较。

示例。通常，当家庭出售其住宅时，收到的房款等于原来的索价吗？通过对 10 个家庭的数据应用 Wilcoxon 带符号等级检验，您可以了解到：7 个家庭收到的房款低于索价，1 个家庭收到的房款高于索价，而 2 个家庭收到的房款等于索价。

统计。平均值、标准差、最小值、最大值、非缺失个案数和四分位数。检验：Wilcoxon 带符号等级检验、符号和 McNemar。如果安装了 Exact Tests 选项（仅在 Microsoft Windows 操作系统上可用），也可使用边际同质性检验。

两个相关样本检验数据注意事项

数据。使用可以排序的数值变量。

假设。虽然没有为两个变量假定特定的分布，但假定配对差分的总体分布是对称的。

获取两个相关样本检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > 2 个相关样本...
2. 选择一个或多对变量。

两个相关样本检验：类型

本节中的检验比较两个相关变量的分布。要使用的适当检验取决于数据类型。

如果数据是连续的，可使用符号检验或 Wilcoxon 带符号等级检验。**符号检验**计算所有个案的两个变量之间的差，并将差分类为正、负或平。如果两个变量分布相似，那么正差和负差的数目不会有很大的差别。**Wilcoxon 带符号等级检验**考虑关于各对之间的差的符号和差的量级的信息。由于 Wilcoxon 带符号等级检验纳入了有关数据的更多信息，因此它比符号检验更为强大。

如果数据为二值数据，那么使用 **McNemar 检验**。此检验通常用于重复测量情况，在此情况中，每个主体的反应将被引出两次，一次在指定事件发生之前，一次在之后。McNemar 检验确定初始响应率（事件前）是否等于最终响应率（事件后）。此检验对于在前后对比设计中检测由实验干预引起的响应变化很有用。

如果数据为分类数据，那么使用**边际同质性检验**。此检验是 McNemar 检验从二值响应到多项式响应的扩展。它检验响应中的变化（使用卡方分布），对于在前后对比设计中检测因实验干预所导致的响应变化很有用。只有安装了 Exact Tests 后，才可用边际同质性检验。

两个相关样本检验: 选项

统计。可以选择一个或全部两个汇总统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。
- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

缺失值。控制对缺失值的处理。

- **按检验排除个案**。如果指定多个检验，将分别独立计算每个检验中的缺失值。
- **按列表排除个案**。从所有分析中排除任何变量具有缺失值的个案。

NPAR TESTS 命令的附加功能（两个相关样本）

使用命令语句语言还可以用列表中的每个变量来检验变量。

请参阅*命令语法参考*以获取完整的语法信息。

多个独立样本检验

“多个独立样本检验”过程比较一个变量上的两组个案或更多组个案。

示例。三种品牌的 100 瓦灯泡，其平均使用寿命不同吗？从 Kruskal-Wallis 单向方差检验分析中，您可能了解到三种品牌灯泡的平均使用寿命的确不同。

统计。平均值、标准差、最小值、最大值、非缺失个案数和四分位数。检验: Kruskal-Wallis *H* 和中位数。

“多个独立样本检验”数据注意事项

数据。使用可以排序的数值变量。

假设。使用独立的随机样本。Kruskal-Wallis *H* 检验要求被检验的样本具有相似的形状。

获取多个独立样本检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > K 个独立样本...

2. 选择一个或多个数值变量。
3. 选择一个分组变量并单击**定义范围**以指定分组变量的最小和最大整数值。

多个独立样本检验: 检验类型

可使用三种检验确定多个独立样本是否来自同一个总体。Kruskal-Wallis *H* 检验、中位数检验和 Jonckheere-Terpstra 检验都检验多个独立样本是否来自同一个总体。

Kruskal-Wallis H 检验是 Mann-Whitney *U* 检验的扩展，它是单向方差检验分析的非参数模拟，用于检测分布位置的差别。**中位数检验**更为通用（但性能有所下降），用于检测位置和形状的分布差别。Kruskal-Wallis *H* 检验和中位数检验假设从其中抽取样本的 *k* 个总体中未进行*先验* 排序。

在 *k* 个总体*已进行* 自然*先验* 排序（升序或降序）的情况下，**Jonckheere-Terpstra 检验**性能更优。例如，*k* 个总体可能代表 *k* 个上升的温度。“不同的温度产生相同的响应分布”这一假设是针对“温度升高，那么响应的量级增加”这一选择进行检验的。此处备选假设已排序，因此，Jonckheere-Terpstra 是最适用的检验。只有安装了“精确检验”附加模块，Jonckheere-Terpstra 检验才可用。

多个独立样本检验：定义范围

要定义范围，请输入对应于分组变量的最低和最高类别的**最小值**和**最大值**的整数值。将排除值在边界以外的个案。例如，如果您指定最小值为 1，最大值为 3，那么只使用 1 到 3 的整数值。最小值必须小于最大值，且必须指定这两个值。

多个独立样本检验：选项

统计。可以选择一个或全部两个汇总统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。
- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

缺失值。控制对缺失值的处理。

- **按检验排除个案**。如果指定多个检验，将分别独立计算每个检验中的缺失值。
- **按列表排除个案**。从所有分析中排除任何变量具有缺失值的个案。

NPAR TESTS 命令的附加功能（K 个独立样本）

使用命令语言还可以为中位数检验指定不同于观察到的中位数的值（使用 MEDIAN 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

多个关联样本检验

“多个关联样本检验”过程比较两个或更多变量的分布情况。

示例。医生、律师、警官和教师在公众中享有不同的声望吗？要求十个人对这四种职业的声望进行排序。Friedman 检验指示这四种职业的确在公众中享有不同的声望。

统计。平均值、标准差、最小值、最大值、非缺失个案数和四分位数。检验：Friedman，Kendall *W* 以及 Cochran *Q*。

“多个关联样本检验”数据注意事项

数据。使用可以排序的数值变量。

假设。非参数检验不要求假定基础分布的形状。而是使用从属的随机样本。

获取多个关联样本检验

1. 从菜单中选择:

   分析 > 非参数检验 > 传统对话框 > K 个相关样本...

2. 选择两个或更多数值检验变量。

多个关联样本检验：检验类型

有三种检验可用于比较多个相关变量的分布。

Friedman 检验是以下两项的非参数等同检验：单样本重复测量设计，或者每个单元格一个观察值的二阶方差分析。Friedman 检验 *k* 个相关变量来自同一总体的原假设。对于每个个案，*k* 个变量的等级从 1 到 *k*。检验统计基于这些等级。

Kendall 的 W 是 Friedman 统计的标准化形式。Kendall 的 *W* 可解释为协调系数，它是评分者之间一致程度的测量。每个个案是一名裁判员或评分者，每个变量是被裁判的一项或一个人。对于每个变量，要计算等级之和。Kendall 的 *W* 的范围从 0（完全不一致）到 1（完全一致）。

Cochrans Q 等于 Friedman 检验，但它适用于所有响应都是二元响应的情况。该检验是 McNemar 检验对 *k* 样本情况的扩展。Cochrans *Q* 检验多个相关二分变量具有相同平均值的假设。对相同的个体或匹配的个体测量变量。

多个关联样本检验：统计

您可以选择统计。

- **描述性**。显示平均值、标准差、最小值、最大值和非缺失个案数。
- **四分位数**。显示对应于第 25 个、第 50 个和第 75 个百分位数的值。

NPAR TESTS 命令的附加功能（K 个相关样本）

请参阅*命令语法参考*以获取完整的语法信息。

第 28 章 多重响应分析

多重响应分析

有两个过程可以用于分析多二分集和多类别集。“多响应频率”过程显示频率表。“多响应交叉表”过程显示二维和三维交叉制表。使用任一过程前，都必须先定义多响应集。

示例。本示例说明多重响应项在市场研究调查中的使用。数据是虚构的，不能解释为实际情况。一个航空公司可能会对飞某条航线的乘客作一个调查，从而评估与其竞争的航空公司。本示例中，美国航空公司要知道乘客在芝加哥到纽约这条航线上乘坐其他航空公司的飞机的情况，并了解航班时间和服务在选择航空公司时的相对重要性。乘务员在每个乘客登机时发给他们一张简单的问卷。第一个问题是: 请选择过去六个月内您在这条航线上至少搭乘过一次的所有航空公司，并在上面画圈 -- 美国航空公司、美联航、TWA、USAir 以及其他。这是一个多重响应问题，因为乘客可以选择多个答案。但是，这个问题不能直接编码，因为一个变量对应每个个案只能有一个值。您必须使用多个变量来映射每个问题的响应。有两种方法可以解决这一问题。一种是对应每个选项定义一个变量（例如，美国航空公司、美联航、TWA、USAir 和其他）。如果乘客选择“美国联合航空公司”，那么指定*美联航*这个变量的代码为 1，如果不选择就是 0。这是映射变量的**多二分法**。另一个映射响应的方法是**多类别法**。用这种方法，您先估计对问题的可能响应的最大数量，然后设置相同数量的变量，用代码指定所乘坐的航空公司。在研究了问卷的样本后，您可能会发现，在过去的六个月中没有乘客在这条航线上坐过三个以上航空公司的飞机。而且，您发现由于对航线没有限制，在“其他”类中，乘客写出了 10 条其他航线。如果使用多重响应法，您将定义三个变量，每个变量的编码为 1 = *美国航空公司*、2 = *美联航*、3 = *twa*、4 = *usair*、5 = *delta* 等等。如果某个乘客选择了美国航空公司和 TWA，那么第一个变量的代码为 1，第二个变量的代码为 3，而第三个变量的代码则是缺失的。另一个乘客可能选择了美国航空公司并在“其他”类写了 Delta。这样，第一个变量的代码为 1，第二个变量的代码为 5，而第三个变量的代码是缺失的。而如果使用多二分法，结果就会得到 14 个独立的变量。虽然在这个调查中两种映射方法都可以使用，但选择哪一个取决于响应的分布。

多重响应定义集

定义多响应集过程将基本变量分组为多二分集和多类别集，您可以获得这些集的频率表和交叉制表。可以定义多达 20 个的多响应集。每个集必须有一个唯一的名称。要删除一个集，请在多响应集列表中将其突出显示，然后单击**删除**。要更改一个集，请在列表中将其突出显示，修改任意集定义特征，然后单击**更改**。

可以将基本变量编码为二分或分类变量。要使用二分变量，请选择**二分**来创建多二分集。对“已计算的值”值输入一个整数值。每个至少出现过一个计数值的变量成为多重二分集的一个类别。选择**类别**来创建与成分变量有相同值范围的多类别集。输入多类别集的类别范围的最小和最大整数值。此过程会将所有成分变量包含范围内的每个不同整数相加求和。空类别不进行制表。

每个多响应集必须指定一个唯一的名称，名称最多可以有七个字符。此过程在您指定的名称前加上美元符号 ($)。不能使用以下保留名称: *casenum*、*sysmis*、*jdate*、*date*、*time*、*length* 和 *width*。多响应集的名称仅在用于多重响应过程时存在。在其他过程中不能使用多响应集名称。另外还可以输入多响应集的描述性变量标签。标签最长可以有 40 个字符。

定义多响应集

1. 从菜单中选择:

   分析 > 多重响应 > 定义变量集...

© Copyright IBM Corporation 1989, 2014

2. 选择两个或更多个变量。
3. 如果变量以二分法进行编码，那么请指明要计数的值。如果您的变量编码为分类变量，请定义类别的范围。
4. 为每个多响应集输入唯一名称。
5. 单击**添加**将多响应集添加到定义的集的列表中。

多响应频率

多重响应频率过程生成多响应集的频率表。必须先定义一个或多个响应集（请参阅“多响应定义集”）。

对于多二分集，显示在输出中的类别名称来自为组中的基本变量定义的变量标签。如果没有定义变量标签，就将变量名称用作标签。对于多类别集，类别标签来自组中的第一个变量的值标签。如果第一个变量缺失的类别存在于组中的其他变量中，请为缺失的类别定义一个值标签。

缺失值。按逐个表的方式排除带缺失值的个案。或者，你可以选择以下两个方式中的一个或全部:

- **排除二分中的个案列表情况**。从多二分集的制表中排除具有任何变量的缺失值的个案。该项仅应用于定义为二分变量的多响应集。缺省的情况下，如果多二分集中的某个个案的成分变量没有一个包含计数的值，就认为该个案缺失。只要至少一个变量包含计数值，那么即使个案中有一些（但不是全部）变量的值缺失，这些个案也包括在组的制表中。
- **排除类别内的个案列表情况**。从多类别集的制表中排除具有任何变量的缺失值的个案。这仅应用于定义为类别集的多响应集。缺省的情况下，对于多类别集，仅当某个个案的成分没有一个包含定义范围内的有效值时，才认为该个案缺失。

示例。每个从调查问题创建的变量都是基本变量。要分析多重响应项，必须将变量合并为两类多响应集中的一类: 多二分集或多类别集。例如，如果一家航空公司调查在过去的六个月中您乘坐了三家航空公司（美国航空公司，美联航，TWA）中的哪家，您使用了二分变量，并定义了**多二分集**，那么该集中的三个变量的每一个都将变为组变量中的一个类别。三家航空公司的计数和百分比显示在一个频率表中。如果发现没有响应者提到两家以上的航空公司，就可以创建两个变量，每个变量有三个代码，每家航空公司一个代码。如果定义**多类别集**，那么通过将基本变量中相同的代码添加到一起来制作值的图表。结果得到的值的集与每个基本变量的集相同。例如，30 个“美联航”的回答是航线 1 的 5 个“美联航”的回答加上航线 2 的 25 个“美联航”的回答。三条航线的计数和百分比显示在一个频率表中。

统计。频率表显示计数、响应百分比、个案百分比、有效个案数目和缺失个案的数量。

多重响应频率数据注意事项

数据。使用多响应集。

假设。计数和百分比提供来自任何分布的数据的有用说明。

相关过程。使用多重响应定义集过程可以定义多响应集。

要获得多重响应频率

1. 从菜单中选择:

   分析 > 多重响应 > 频率...

2. 选择一个或多个多响应集。

多响应交叉表

“多响应交叉表”过程对定义的多响应集、基本变量或组合进行交叉制表。还可以获得基于个案或响应的单元格百分比，修改缺失值的处理方式，或者获取成对的交叉制表。必须先定义一个或多个响应集（请参阅“定义多响应集”）。

对于多二分集，显示在输出中的类别名称来自为组中的基本变量定义的变量标签。如果没有定义变量标签，就将变量名称用作标签。对于多类别集，类别标签来自组中的第一个变量的值标签。如果第一个变量缺失的类别存在于组中的其他变量中，请为缺失的类别定义一个值标签。此过程可显示三行中列的类别标签，每一行最多可以有八个字符。为了避免断开单词，可以互换行和列项，或者重新定义标签。

示例。多二分集和多类别集都可以与本过程中的其他变量进行交叉制表。一间航空公司的乘客调查请乘客提供以下信息：在下列选项中选择您在过去六个月内至少搭乘过一次的航空公司（美国航空公司、美联航和TWA）。在选择航班时间或服务时，哪个更重要？请只选择一个答案。在将数据作为二分变量或多分类变量输入并将它们组合为一个集后，您就可以将航线选择与涉及服务或航班时间的问题进行交叉制表了。

统计。将单元格、行、列和总计数与单元格、行、列和总百分比进行交叉制表。单元格百分比可以基于个案或响应。

多重响应交叉表数据注意事项

数据。使用多响应集或数字分类变量。

假设。计数和百分比提供来自任意分布的数据的有用说明。

相关过程。使用多重响应定义集过程可以定义多响应集。

要获得多重响应交叉表

1. 从菜单中选择:

   分析 > 多重响应 > 交叉表...

2. 为每个交叉制表的维度选择一个或多个数值变量或多响应集。
3. 定义每个基本变量的范围。

还可以获得控制变量或多响应集的每个类别的二阶交叉制表。为“层”列表选择一项或多项。

多重响应交叉表: 定义范围

必须定义交叉制表中的任何基本变量的值范围。输入要制表的最小和最大类别整数值。该范围外的类别将不包括在分析中。包含范围内的值被认为是整数（非整数的小数部分将被截去）。

多重响应交叉表: 选项

单元格百分比。始终显示单元格计数。可以选择显示行百分比、列百分比和双向表（总）百分比。

百分比基于。单元格百分比可基于个案（或响应者）。如果选择跨响应集匹配变量，那么该选项不可用。单元格百分比也可以基于响应。对于多二分集，响应的数量等于个案中已计算的值。对于多类别集，响应的数量等于位于所定义范围内的值的数量。

缺失值。可以选择以下方式中的一个或两个都选:

- **排除二分中的个案列表情况**。从多二分集的制表中排除具有任何变量的缺失值的个案。该项仅应用于定义为二分变量的多响应集。缺省的情况下，如果多二分集中的某个个案的成分变量没有一个包含计数的值，就认为该个案缺失。只要至少一个变量包含计数值，那么即使个案中有一些（但不是全部）变量的值缺失，这些个案也包括在组的制表中。
- **排除类别内的个案列表情况**。从多类别集的制表中排除具有任何变量的缺失值的个案。这仅应用于定义为类别集的多响应集。缺省的情况下，对于多类别集，仅当某个个案的成分没有一个包含定义范围内的有效值时，才认为该个案缺失。

缺省的情况下，在将两个多类别集进行交叉制表时，此过程对第一组中的每个变量与第二组中的每个变量进行制表，并计算每个单元格的计数的总和；因此，有些响应在表中可能会多次出现。可以选择以下选项:

跨响应集匹配变量。将第一组中的第一个变量与第二组中的第一个变量配对，依此类推。如果选择该选项，过程将单元格百分比基于响应，而不是响应者。对于多二分集或基本变量，配对操作不可用。

MULT RESPONSE 命令附加功能

使用命令语法语言还可以:

- 获得多达五个维度的交叉制表（用 BY 子命令）。
- 更改输出格式选项，包括不显示值标签（用 FORMAT 子命令）。

请参阅*命令语法参考*以获取完整的语法信息。

第 29 章 报告结果

报告结果

个案列表和描述统计是研究和显示数据的基本工具。可以通过数据编辑器或“摘要”过程获取个案列表，通过“频率”过程获取频率计数和描述统计，通过“平均值”过程获取子群体统计。以上每个都使用了为使信息更清晰而设计的格式。如果想要以不同的格式显示信息，可以使用“按行汇总”和“按列汇总”根据需要控制数据显示。

按行汇总

“按行汇总”生成的报告中，不同的汇总统计按行显示。还提供带有或不带汇总统计的个案列表。

示例。有多家连锁零售店的公司保存员工信息的记录，包括薪金、工作期和每名员工所工作的商店和部门。您可以生成一份报告，该报告提供个别的雇员信息（列表），这些信息按商店和部门分组（分组变量），提供有每家商店、部门和每家商店内部的部门的汇总统计（例如平均薪金）。

数据列。列出需要个案列表或汇总统计的报告变量，并控制数据列的显示格式。

中断列。列出将报告划分为各个组的可选分组变量，并控制汇总统计以及中断列的显示格式。对于多个分组变量，对列表中先前分组变量的类别中的每个分组变量的每个类别均有一个单独的组。分组变量应是离散型分类变量，它们将个案划分为有限个有具体含义的类别。每个分组变量的个别值排好序之后，作为单独一列出现在所有数据列的左边。

报告。控制报告的整体特征，包括整体汇总统计、缺失值的显示、页编号和标题。

显示个案。为每个个案显示数据列变量的实际值（或值标签）。这会生成列表报告，该报告可能比摘要报告要长得多。

预览。只显示报告的第一页。此选项对于预览报告的格式但不处理整个报告很有用。

已经对数据排序。对于具有分组变量的报告，在生成报告之前必须按分组变量值对数据文件排序。如果数据文件已经按分组变量的值排序，那么可以通过选择此选项来节省处理时间。在运行预览报告之后，此选项尤其有用。

获取摘要报告: 按行汇总

1. 从菜单中选择:

   分析 > 报告 > 按行汇总...
2. 选择一个或多个变量作为“数据列”。对于每个选定的变量，在报告中均生成一列。
3. 对于已经按子组排序并显示的报告，选择一个或多个变量作为“中断列”。
4. 对于带有分组变量定义的子组的汇总统计的报告，在“中断列变量”列表中选择分组变量并单击“中断列”组中的**摘要**指定汇总测量。
5. 对于带有整体汇总统计的报告，单击**摘要**指定汇总测量。

© Copyright IBM Corporation 1989, 2014

报告数据列/中断格式

“格式”对话框控制列标题、列宽、文本对齐方式以及数据值或值标签的显示。“数据列格式”控制报告页右侧数据列的格式。“中断格式”控制左侧中断列的格式。

列标题。对于选定的变量，控制列标题。长标题在列中自动换行。使用 Enter 键可手工在想要标题换行的地方插入换行符。

列中位数的位置。对于选定的变量，控制列中数据值或值标签的对齐方式。值或标签的对齐方式不影响列标题的对齐方式。您可以使列内容缩进指定的字符数，或者使内容居中。

列内容。对于选定变量，控制数据值或定义的值标签的显示。对于没有定义的值标签的任何值，始终显示数据值。（不适用于列摘要报告中的数据列。）

报告: 摘要行/最终摘要行

这两个“摘要行”对话框控制分类组的汇总统计以及整个报告的汇总统计的显示。“摘要行”控制由分组变量定义的每个类别的子组统计。“最终摘要行”控制显示在报告末尾的整体统计。

可用的汇总统计有值的和、值的平均值、最小值、最大值、个案号、指定值以上或以下的个案百分比、指定值范围内的个案百分比、标准差、峰度、方差和偏度。

报告: 中断选项

“中断选项”控制中断类别信息的间距和页编号。

页面控制。为选定分组变量的类别控制间距和分页。您可以在中断类别之间指定多个空白行，或者在新的页面上开始每个中断类别。

摘要前的空行。控制中断类别标签之间，或者数据与汇总统计之间的空白行数。对于既包含个别个案列表又包含中断类别的汇总统计的组合报告，这尤其有用；在这些报告中，您可以在个案列表之间和汇总统计之间插入空格。

报告: 选项

“报告: 选项”控制缺失值和报告页编号的处理和显示。

按列表排除含有缺失值的个案。（从报告中）排除任何报告变量具有缺失值的所有个案。

缺失值显示为。允许您指定在数据文件中代表缺失值的符号。该符号只能是一个字符，用于表示*系统缺失*值和*用户缺失*值。

计算页数的起点。允许您为报告的第一页指定页码。

报告: 布局

“报告: 布局”控制每个报告页的宽度和长度、页面上报告的放置以及空白行和标签的插入。

页面布局。控制以行数（顶部和底部）和字符数（左和右）表示的页边距以及页边距内的报告对齐方式。

页标题和页脚。控制将页标题和页脚与报告正文分隔开来的行数。

中断列。控制中断列的显示。如果指定了多个分组变量，那么它们可以位于分开的列或位于第一列。将所有分组变量放置在第一列会生成较窄的报告。

列标题。控制列标题的显示，包括标题下划线、标题和报告正文之间的间隔以及列标题的垂直对齐方式。

数据列行与分组标注。控制数据列信息（数据值和/或汇总统计）相对于每个中断类别开始处的中断标签的放置。数据列信息的第一行可以开始于与中断类别标签相同的行，也可以从中断类别标签之后指定的行数处开始。（不适用于列摘要报告。）

报告: 标题

“报告: 标题”控制报告标题和页脚的内容和放置。最多可指定 10 行页面标题和 10 行页脚，并带有可各行为左对齐、居中和右对齐的组件。

如果将变量插入到标题或页脚中，那么变量的当前值标签或值会显示在标题或页脚中。在标题中，会显示与页开始处的变量的值对应的值标签。在页脚中，会显示与页结束处的变量的值对应的值标签。如果没有值标签，那么会显示实际值。

特殊变量。特殊变量 *DATE* 和 *PAGE* 允许您将当前日期或页码插入到报告页眉或页脚的任意一行。如果数据文件包含名为 *DATE* 或 *PAGE* 的变量，那么不能在报告标题或页脚中使用这些变量。

按列汇总

“按列汇总”生成一些摘要报告，在这些报告中，不同的汇总统计显示在单独的列中。

示例。有多家连锁零售店的公司保存关于员工信息的记录，包括薪金、工作期和每名员工所工作的部门。您可以生成一份报告，该报告提供每个部门的摘要薪金统计（例如平均值、最小值和最大值）。

数据列。列出您需要其汇总统计的报告变量，并控制显示格式以及为每个变量显示的汇总统计。

中断列。列出将报告划分为各个组的可选分组变量，并控制中断列的显示格式。对于多个分组变量，对列表中先前分组变量的类别中的每个分组变量的每个类别均有一个单独的组。分组变量应是离散型分类变量，它们将个案划分为有限个有具体含义的类别。

报告。控制报告的整体特征，包括缺失值的显示、页编号和标题。

预览。只显示报告的第一页。此选项对于预览报告的格式但不处理整个报告很有用。

已经对数据排序。对于具有分组变量的报告，在生成报告之前必须按分组变量值对数据文件排序。如果数据文件已经按分组变量的值排序，那么可以通过选择此选项来节省处理时间。在运行预览报告之后，此选项尤其有用。

获取摘要报告: 按列汇总

1. 从菜单中选择:

   分析 > 报告 > 按列汇总...

2. 选择一个或多个变量作为“数据列”。对于每个选定的变量，在报告中均生成一列。
3. 要更改某个变量的汇总测量，从“数据列变量”列表中选择该变量，然后单击**摘要**。
4. 要获取某个变量的多个汇总测量，从源列表中选择该变量并将其多次移至“数据列变量”列表中，为每个需要的汇总测量移动一次。

5. 要显示包含合计、平均值、比率或其他现有列的函数，单击**插入总计**。这会将一个名为*总计*的变量放置到“数据列”列表中。
6. 对于已经按子组排序并显示的报告，选择一个或多个变量作为“中断列”。

数据列汇总函数

“摘要行”控制为选定数据列变量显示的汇总统计。

可用的汇总统计有合计、平均值、最小值、最大值、个案数、指定值以上或以下的个案百分比、指定值范围内的个案百分比、标准差、方差、峰度和偏度。

总计列的数据列摘要

“摘要列”控制对两个或更多数据列进行摘要的总计汇总统计。

可用的总计汇总统计有列的和、列的平均值、最小值、最大值、两列值之间的差分、一列中的值除以另一列中的值的商以及列值相乘的积。

列的和。*总计*列是“摘要列”列表中列的和。

列的平均值。*总计*列是“摘要列”列表中列的平均值。

列的最小值。*总计*列是“摘要列”列表中列的最小值。

列的最大值。*总计*列是“摘要列”列表中列的最大值。

第一列 - 第二列。*总计*列是“摘要列”列表中列的差。“摘要列”列表必须正好包含两列。

第一列/第二列。*总计*列是“摘要列”列表中列的商。“摘要列”列表必须正好包含两列。

% 第一列/第二列。*总计*列是“摘要列”列表中第一列占第二列的百分比。“摘要列”列表必须正好包含两列。

列的积。*总计*列是“摘要列”列表中列的乘积。

报告: 列格式

“按列汇总”的数据列和中断列格式编排选项与“按行汇总”中所描述的相同。

按列汇总: 中断选项

“中断选项”控制中断类别的小计显示、间距以及页编号。

小计。控制中断类别的显示小计。

页面控制。为选定分组变量的类别控制间距和分页。您可以在中断类别之间指定多个空白行，或者在新的页面上开始每个中断类别。

小计前的空行数。控制中断类别数据和小计之间的空白行数。

按列汇总: 选项

这些选项控制总计的显示、缺失值的显示以及列摘要报告中的页编号。

总计。显示并标记每一列的总计；显示在列的底部。

缺失值。您可以将缺失值从报告中排除，或者选择单个字符以在报告中指示缺失值。

列摘要的报告布局

“按列汇总”的报告布局选项与那些为“按行汇总”所描述的报告布局选项相同。

REPORT 命令的附加功能

使用命令语法语言还可以:

- 在单个摘要行的列中显示不同的汇总函数。
- 将摘要行插入到非数据列变量的数据列中，或者插入到汇总函数的各种组合（复合函数）的数据列中。
- 使用“中位数”、“众数”、“频率”和“百分比”作为汇总函数。
- 更精确地控制汇总统计的显示格式。
- 在报告中的各个不同的点插入空白行。
- 在列表报告中每 *n* 个个案后插入空白行。

由于 REPORT 语法的复杂性，您可能会发现，在使用语法构建新报告时，它对于以下操作非常有用: 估计从对话框生成的报告，复制并粘贴相应的语法，以及改进语法以生成与您的需要精确一致的报告。

请参阅*命令语法参考*以获取完整的语法信息。

第 30 章 可靠性分析

可靠性分析允许您研究测量刻度的属性以及组成这些标度的项。“可靠性分析”过程计算标度可靠性的众多常用度量，还提供关于标度中的各项之间关系的信息。类内相关系数可用来计算评分者间的可靠性估计。

示例。我的调查表能以有用的方式度量客户满意度吗？使用可靠性分析，您可以确定调查表中各项的相互关联程度，可以获取重复性的总体指标或作为一个整体的标度的内部一致性，并且可以识别应从标度中排除的问题项。

统计。每个变量和标度的描述、跨项的汇总统计、项之间的相关性和协方差、可靠性估计、ANOVA 表、类内相关系数、Hotelling T^2 以及 Tukey 的可加性检验。

模型。以下可靠性模型可用:

- **Alpha (Cronbach)**。此模型是内部一致性模型，基于平均的项之间的相关性。
- **半分**。此模型将标度拆分成两个部分，并检查两部分之间的相关性。
- **Guttman**。此模型计算 Guttman 的下界以获取真实可靠性。
- **平行**。此模型假设所有项具有相等的方差，并且重复项之间具有相等的误差方差。
- **严格平行**。此模型假设为平行模型，还假设所有项具有相等的平均值。

可靠性分析数据注意事项

数据。数据可以是二分数据、有序数据或区间数据，但数据应是用数值编码的。

假设。观察值应是独立的，且项与项之间的误差应是不相关的。每对项应具有二元正态分布。标度应是可加的，以便每一项都与总得分线性相关。

相关过程。如果想要探索标度项的维数（以查明是否需要多个结构来代表项得分的模式），那么使用因子分析或多维刻度。要标识同类变量组，可使用系统聚类分析以使变量聚类。

获取可靠性分析

1. 从菜单中选择:

   分析 > 刻度 > 可靠性分析...
2. 选择两个或更多变量作为可加标度的可能成分。
3. 从“模型”下拉列表中选择模型。

可靠性分析统计

可以选择多个不同的统计来描述标度和项。缺省报告的统计包括个案数、项数和可靠性估计，如下所示:

- **Alpha 模型**。系数 alpha；对于二分数据，它等同于 Kuder-Richardson 20 (KR20) 系数。
- **半分模型**。形式之间的相关性、Guttman 半分可靠性、Spearman-Brown 可靠性（相等长度和不相等长度）以及每一半的 alpha 系数。
- **Guttman 模型**。可靠性系数 lambda 1 到 lambda 6。
- **平行和严格平行模型**。模型拟合度优度检验；误差方差的估计值、公共方差和真实方差；估计的公共项间相关性；估计的可靠性以及可靠性的无偏估计。

© Copyright IBM Corporation 1989, 2014

描述性。为跨个案的标度或项生成描述统计。

- **项**。为跨个案的项生成描述统计。
- **刻度**。为标度生成描述统计。
- **标度（如果项已删除）**。显示将每一项与由其他项组成的标度进行比较时的汇总统计。这些统计包括：该项从标度中删除时的标度平均值和方差、该项与由其他项组成的标度之间的相关性，以及该项从标度中删除时的 Cronbach alpha 值。

摘要。提供跨标度中所有项的项分布的描述统计。

- *平均值 (Means)*. 项平均值的汇总统计。显示项平均值的最小、最大和平均值，项平均值的范围和方差，以及最大项平均值与最小项平均值的比。
- *方差*。项方差的汇总统计。显示项方差的最小、最大和平均值，项方差的范围和方差，以及最大项方差与最小项方差的比。
- *协方差*。项间协方差的汇总统计。显示项之间的协方差的最小、最大和平均值，项之间的协方差的范围和方差，以及最大项之间协方差与最小项之间的协方差的比。
- *相关性*。项之间的相关性的汇总统计。显示项之间的相关性的最小、最大和平均值，项间相关性的范围和方差，以及最大项之间的相关性与最小项之间的相关性的比。

项之间。生成项与项之间的相关性矩阵或协方差矩阵。

ANOVA 表。生成相等平均值的检验。

- *F 检验*。显示重复测量方差分析表。
- *Friedman 卡方*。显示 Friedman 的卡方 Kendall 的协同系数。此选项适用于以等级为形式的数据。卡方检验在 ANOVA 表中替换通常的 F 检验。
- *Cochran 卡方*。显示 Cochrans Q。此选项适用于双分支。Q 统计在 ANOVA 表中替换通常的 F 统计。

Hotelling T 平方。生成以下原假设的多变量检验：标度上的所有项具有相同的平均值。

Tukey 的可加性检验。生成以下假设的检验：项中不存在可乘交互关系。

类内相关系数。生成个案内值的一致性或符合度的测量。

- **模型**。选择用于计算类内相关系数的模型。可用的模型为双向混合、双向随机和单向随机。当人为影响是随机的，而项的作用固定时，选择**双向混合**；当人为影响和项的作用均为随机时选择**双向随机**。当人为影响随机时选择**单向随机**。
- **类型**。选择指标类型。可用的类型为“一致”和“绝对一致”。
- **置信区间**。指定置信区间的置信度。缺省值为 95%。
- **检验值**。指定假设检验系数的假设值。该值是用来与观察值进行比较的值。缺省值为 0。

RELIABILITY 命令的附加功能

使用命令语法语言还可以:

- 读取并分析相关性矩阵。
- 写相关性矩阵以用于以后的分析。
- 指定与半分法的相等半分不同的拆分方法。

请参阅*命令语法参考*以获取完整的语法信息。

第 31 章 多维刻度

多维刻度尝试寻找对象间或个案间一组距离测量的结构。该任务是通过将观察值分配到概念空间（通常为二维或三维）中的特定位置实现的，这样使空间中的点之间的距离尽可能与给定的非相似性相匹配。在很多情况下，这个概念空间的维度可以解释并可以用来进一步分析数据。

如果您已经客观地度量了变量，就可以将多维刻度用作减少数据的技巧（在必要时，“多维刻度”过程可根据多变量数据计算距离）。还可使用多维刻度进行对象间或概念间的非相似性的主观评定。此外，“多维刻度”过程还可以处理来自多个源的非相似性数据，因为可能会有多个评分者或问卷响应者。

示例。人们如何理解不同汽车之间的关系？如果响应者的数据表明对不同构造和车型的汽车有相似性等级评定，您就可以使用多维刻度来描述消费者观点的维度。例如，您可能会发现交通工具的价格和大小构成一个二维空间，该空间就代表响应者所反映的相似性。

统计。对于每个模型：数据矩阵、最优刻度化数据矩阵、S 应力 (Young)、应力 (Kruskal)、RSQ、刺激坐标、平均应力以及每项刺激的 RSQ（RMDS 模型）。对于各个差异 (INDSCAL) 模型：每个主体的主体权重和古怪指数。对于复制的多维刻度模型中的每个矩阵：每项刺激的应力和 RSQ。图：刺激坐标（二维或三维），不同点与距离的散点图。

多维刻度数据注意事项

数据。如果数据是非相似性数据，那么所有的非相似性都应该是定量的，应该用相同的刻度进行度量。如果数据是多变量数据，变量可以为定量数据、二分类数据或计数数据。变量刻度是一个重要问题 -- 刻度之间的差异可能会影响解。如果变量在刻度上有很大差异（例如：一个变量以美元为单位度量，而另一个以年数为单位度量），那么应该考虑对它们进行标准化（这可以通过多维刻度过程来自动完成）。

假设。相对来讲，多维刻度过程没有分布假定。请确保在“多维刻度: 选项”对话框中选择适当的测量级别（有序、定距或定比）以确保正确计算结果。

相关过程。如果您的目标是减少数据，特别当变量为定量时，就可以考虑使用另一种方法，即因子分析。如果要确定相似个案的组，请考虑使用分层聚类或 *k*-means 聚类分析补充多维刻度分析。

获得多维刻度分析

1. 从菜单中选择:

   分析 > 刻度 > 多维刻度...
2. 选择至少四个数值变量用于分析。
3. 在“距离”组中，选择**数据为距离数据**或**从数据创建距离**。
4. 如果您选择**从数据创建距离**，也可以为单个矩阵选择分组变量。分组变量可为数值或字符串。

根据需要，您可以:

- 指定当数据为距离数据时距离矩阵的形状。
- 指定在从数据创建距离时使用的距离测量。

© Copyright IBM Corporation 1989, 2014

多维刻度: 数据形状

如果您的活动数据集代表一组对象中的距离或者代表两组对象之间的距离，那么指定数据矩阵的形状才能得到正确的结果。

注: 如果“模型”对话框指定行条件性，那么不能选择**正对称**。

多维刻度: 创建测量

多维刻度使用非相似性数据创建刻度分析解。如果您的数据为多变量数据（度量到的变量的值），就必须创建非相似性数据才能计算多维刻度解。可以指定从数据创建非相似性测量的详细信息。

度量。允许您指定进行分析的非相似性测量。从与您的数据类型相关的“测量”组选择一个选项，然后从与那一类测量相关的下拉列表选择一种测量。可以使用的选项有:

- **区间**。欧氏距离、平方 Euclidean 距离、Chebychev、区组、Minkowski 或定制。
- **计数**。卡方统计测量或 Phi 平方统计测量。
- **二元**。欧氏距离、平方 Euclidean 距离、刻度差分、模式差分、方差或 Lance 和 Williams。

创建距离矩阵。可以选择要分析的单位。选项有“变量之间”或“个案之间”。

转换值。在某些情况下（例如用相差很大的刻度测量变量时），您可能希望先将值标准化，然后再计算近似值（对二分类数据不可用）。从“标准化”下拉列表中选择一个标准化方法。如果不需要标准化，那么选择**无**。

多维刻度: 模型

多维刻度分析模型的正确估计取决于数据的方面和模型本身的方面。

测量级别。使您可以指定数据的级别。选项为“序数”、“区间”或“比率”。如果变量是有序的，选择**打开结观察值**要求将它们当作连续变量，这样结（不同个案的相等值）就会最优化解开。

条件性。可以指定哪些比较是有意义的。选项为“矩阵”、“行”或“无约束”。

维。使您可以指定刻度解决方案的维度性。对该范围中的每个数字都计算出一个答案。指定 1 到 6 之间的整数；只有将**Euclidean 距离**选为刻度模型时才可以使用最小为 1 的值。要获得单一的解，为最小值和最大值指定同一个值。

刻度模型。使您可以指定作为刻度执行标准的假定。可以使用的选项为“Euclidean距离”或“个别差异 Euclidean 距离”（也称为 INDSCAL）。对于“个别差异 Euclidean 距离”模型，如果适用于您的数据的话，就可以选择**允许负的主体权重**。

多维刻度: 选项

可以为多维刻度分析指定选项。

输出。可以选择各种输出类型。可用的选项为“组图”、“个别主体图”、“数据矩阵”以及“模型和选项摘要”。

标准。可以确定迭代停止的时间。要更改缺省值，请输入 **S 应力收敛性**、**最小 S 应力值**和**最大迭代次数**的值。

将小于 n 的距离看作缺失值。从分析中排除小于该值的距离。

ALSCAL 命令附加功能

使用命令语法语言还可以:

- 使用有关多维刻度的文献中的其他三种类型模型，即 ASCAL、AINDS 和 GEMSCAL。
- 对定距数据和定比数据进行多项式的转换。
- 用有序数据分析相似性（而不是距离）。
- 分析名义数据。
- 将各种坐标和权重矩阵保存到文件，并重新读取进行分析。
- 限制多维展开。

请参阅*命令语法参考*以获取完整的语法信息。

第 32 章 比率统计

“比率统计”过程提供了一个描述两个刻度变量间比率的汇总统计的综合列表。

可以升序或降序按分组变量的值对输出排序。可以在输出中隐藏比率统计报表，并将结果保存到外部文件。

示例。五个县的每一个县中，住宅的估价和售价之间的比率是否存在良好的一致性？从输出中，您会了解到各县的比率分布显著不同。

统计。中位数、平均值、加权平均值、置信区间、离差系数 (COD)、以中位数为中心的变异系数、以平均值为中心的变异系数、价格相关微分 (PRD)、标准差、平均绝对偏差 (AAD)、范围、最小和最大值、对用户指定的范围或中位数比率中的百分比所计算的集中指数。

比率统计数据注意事项

数据。使用数值代码或字符串以对分组变量进行编码（名义或序数级别测量）。

假设。定义比率的分子和分母的变量应是取正值的刻度变量。

获取比率统计

1. 从菜单中选择:

   分析 > 描述统计 > 比率...
2. 选择分子变量。
3. 选择分母变量。

或者:

- 选择分组变量并指定结果中组的排序方式。
- 选择是否在“查看器”中显示结果。
- 选择是否将结果保存到外部文件以供以后使用，并指定保存结果的文件的名称。

比率统计

集中趋势。集中趋势的测量是描述比率分布的统计。

- **中位数**。中位数是这样的一个值，小于该值的比率数与大于该值的比率数相等。
- **平均值。**比率的总和除以比率的总数所得到的结果。
- **权重平均值。**分子的平均值除以分母的平均值所得到的结果。加权平均值也是比率按分母加权之后的平均值。
- **置信区间**。显示平均值、中位数和加权平均值（如果要求）的置信区间。可指定大于等于 0 且小于 100 的值作为置信区间。

离差。这些统计测量观察值中的变差量或分散量。

- **AAD**。平均绝对偏差是中位数比率的绝对离差求和并用结果除以比率总数所得的结果。
- **COD**。离差系数是将平均绝对偏差表示为中位数的百分比的结果。
- **PRD**。价格相关微分也称为回归指数，是平均值除以加权平均值得出的结果。

© Copyright IBM Corporation 1989, 2014

- **中位数居中的 COV**。中位数居中的变异系数是将与中位数偏差的均方根表示为中位数百分比的结果。
- **平均值居中的 COV**。平均值居中的变异系数是将标准差表示为平均值百分比的结果。
- **标准差**。标准差是比率与平均值间偏差的平方之和，再除以比率总数减一，取正的平方根所得到的结果。
- **范围**。范围是最大的比率减去最小的比率所得的结果。
- **最小值**。最小值是最小的比率。
- **最大值**。最大值是最大的比率。

集中指数。集中系数度量落在某个区间中的比率的百分比。有两种方法计算该值:

- **在以下比比率之间**。在这里，区间是通过指定区间的最小值和最大值而显式定义的。输入最小比值和最大比值，并单击**添加**可获得区间。
- **在以下比率之内**。在这里，区间是通过指定中位数的百分比而隐式定义的。输入 0 到 100 之间的值并单击**添加**。区间的下界等于 (1 – 0.01 × 值) × 中位数，上界等于 (1 + 0.01 × 值) × 中位数。

第 33 章 ROC 曲线

对于按一个变量的两种类别对主体进行分类的设计，此过程是评估其分类设计性能的有用方法。

示例。银行感兴趣的是正确地将客户分类成会拖欠贷款和不会拖欠贷款两类，因此为做出这些决策制定了特殊的方法。ROC 曲线可用来评估这些方法的效果如何。

统计。ROC 曲线下的区域，以及置信区间和 ROC 曲线的坐标点。图：ROC 曲线。

方法。ROC 曲线下的区域的估计可以非参数方式计算，也可以使用双负指数模型以参数方式计算。

ROC 曲线数据注意事项

数据。检验变量是定量变量。检验变量的组成要素常为由判别分析或 logistic 回归所得的概率，或是某个指示评分者“确信度”（主体落入一个类别或另一个类别的范围内）的随意刻度上的得分。状态变量可以是任何类型，并指示主体真正所属的类别。状态变量的值指示哪一个类别应视为*正的*。

假设。假设评分者刻度上的数字增加表示主体属于一个类别的可信度增加，而刻度上的数字减少表示主体属于另一个类别的可信度增加。用户必须选择哪一个方向是*正的*。还假设已知每个主体*真正*所属的类别。

获取 ROC 曲线

1. 从菜单中选择:

   分析 > ROC 曲线...
2. 选择一个或多个检验概率变量。
3. 选择一个状态变量。
4. 确定状态变量的*正*值。

ROC 曲线: 选项

您可以为 ROC 分析指定以下选项:

分类。允许您指定进行*正*分类时，是包含还是排除分界值。当前此设置对输出没有影响。

检验方向。允许您指定相对于*正*类别的刻度方向。

区域的标准误差的参数。允许您指定估计曲线以下区域的标准误差的方法。可用的方法是非参数法和双负指数法。还允许您设置置信区间的置信度。可用的范围是 50.1% 到 99.9%。

缺失值。允许您指定如何处理缺失值。

第 34 章 模拟

预测模型（例如线性回归）需要一组已知输入来预测结果或目标值。然而，在许多实际应用中，输入值通常是不确定的。模拟允许您考虑预测模型输入的不确定性，并且在存在不确定性的情况下评估各种模型结果的可能性。例如，您有一个包含材料成本作为输入的盈利模型，但由于市场波动在成本上存在不确定性。您可以使用模拟来对此不确定性进行建模，并确定它对利润的影响。

IBM SPSS Statistics 中的模拟使用 Monte Carlo 法。采用概率分布（如三角分布）来对不确定输入进行建模，并从这些分布抽取生成这些输入的模拟值。值已知的输入保持固定为已知值。采用每个不确定输入的模拟值和已知输入的固定值来评估预测模型，以计算模型目标。此过程将重复多次（通常为数万次或数十万次），以取得目标值分布的结果，并可用于回答有关概率特性的疑问。在 IBM SPSS Statistics 上下文中，此过程的每次重复都将生成不同的数据个案（记录），其中包含不确定输入的模拟值集合、固定输入的值以及模型的一个或多个预测目标。

通过对所要模拟的变量指定概率分布，还可以在没有预测模型的情况下模拟数据。生成的每个数据个案都包含所指定变量的一组模拟值。

要运行模拟，您需要指定详细信息，例如预测模型、不确定输入的概率分布、这些输入之间的相关性，以及任何固定输入值。在指定模拟的所有详细信息之后，可以运行该模拟，并可以选择性地将这些指定保存到**模拟计划**文件中。您可以将模拟计划共享给其他用户，然后他们可以在无需了解其创建细节的情况下运行模拟。

可以在两个界面上使用模拟。“模拟构建器”是可供用户设计和运行模拟的高级界面。它提供完整功能集来设计模拟、将规范保存为模拟计划文件、指定输出和运行模拟。您可以构建基于 IBM SPSS 模型文件或者模拟构建器中定义的一组定制方程的模拟。您还可以在“模拟构建器”中加载现有模拟计划，修改相关设置并运行模拟，或者选择保存已更新的模拟计划。对于那些拥有模拟计划且主要打算运行模拟的用户，可以使用一个较简单的界面。它允许您修改相关设置以便在不同条件下运行模拟，但它并未提供用于设计模拟的“模型构建器”完整功能。

基于模型文件设计模拟

1. 从菜单中选择:

   分析 > 模拟...
2. 单击**选择 SPSS 模型文件**，然后单击**继续**。
3. 打开模型文件。

   模型文件是一个 XML 文件，其中包含根据 IBM SPSS Statistics 或 IBM SPSS Modeler 创建的模型 PMML。请参阅 第 152 页的『“模型”选项卡』主题以获取更多信息。
4. 在“模拟构建器”的“模拟”选项卡上，指定模拟输入的概率分布和固定输入值。如果活动数据集包含模拟输入的历史数据，请单击**全部拟合**以自动确定与每个此类输入的数据拟合最紧密的分布，并确定它们之间的相关性。对于每个与历史数据不拟合的模拟输入，您必须通过选择分布类型并输入必需参数明确指定分布。
5. 单击**运行**以运行模拟。缺省情况下，用于指定模拟详细信息的模拟计划将被保存到在“保存”设置上指定的位置。

可用选项有:

© Copyright IBM Corporation 1989, 2014

- 修改保存的模拟计划的位置。
- 指定模拟输入之间的已知相关性。
- 自动计算分类输入之间的关联的列联表，并在为这些输入生成数据时使用这些关联。
- 指定敏感度分析，以调查改变固定输入的值或者改变模拟字段的分布参数所产生的效应。
- 指定高级选项，例如设置要生成的最大个案数或者请求进行尾部取样。
- 自定义输出。
- 将模拟数据保存为数据文件。

基于自定义方程设计模拟

1. 从菜单中选择:

   分析 > 模拟...
2. 单击**输入方程**，然后单击**继续**。
3. 在模型构建器的“模型”选项卡上单击**新建方程**，以便在预测模型中定义每个方程。
4. 单击“模拟”选项卡，并指定模拟输入的概率分布以及固定输入的值。如果活动数据集包含模拟输入的历史数据，请单击**全部拟合**以自动确定与每个此类输入的数据拟合最紧密的分布，并确定它们之间的相关性。对于每个与历史数据不拟合的模拟输入，您必须通过选择分布类型并输入必需参数明确指定分布。
5. 单击**运行**以运行模拟。缺省情况下，用于指定模拟详细信息的模拟计划将被保存到在“保存”设置上指定的位置。

可用选项有:

- 修改保存的模拟计划的位置。
- 指定模拟输入之间的已知相关性。
- 自动计算分类输入之间的关联的列联表，并在为这些输入生成数据时使用这些关联。
- 指定敏感度分析，以调查改变固定输入的值或者改变模拟字段的分布参数所产生的效应。
- 指定高级选项，例如设置要生成的最大个案数或者请求进行尾部取样。
- 自定义输出。
- 将模拟数据保存为数据文件。

在没有预测模型的情况下设计模拟

1. 从菜单中选择:

   分析 > 模拟...
2. 单击**创建模拟数据**，然后单击**继续**。
3. 在模拟构建器中的“模型”选项卡上，选择要模拟的字段。您可以从活动数据集中选择字段，也可以通过单击**新建**来定义新字段。
4. 单击“模拟”选项卡，并指定要模拟的字段的概率分布。如果活动数据集包含其中任何字段的历史数据，请单击**全部拟合**以自动确定与数据拟合最紧密的分布以及确定这些字段之间的相关性。对于与历史数据不拟合的字段，您必须通过选择分布类型并输入必需参数明确指定分布。
5. 单击**运行**以运行模拟。缺省情况下，模拟的数据将保存到“保存”设置中指定的新数据集。另外，还会将模拟计划（用于指定模拟详细信息）保存到“保存”设置中指定的位置。

可用选项有:

- 修改模拟的数据或者保存的模拟计划的位置。
- 指定模拟字段之间的已知相关性。
- 自动计算分类字段之间的关联的列联表，并在为这些字段生成数据时使用这些关联。
- 指定敏感度分析，以调查改变模拟字段的分布参数所产生的效应。
- 指定高级选项，例如设置要生成的个案数。

从模拟计划运行模拟

有两个选项可用于从模拟计划运行模拟。您可以使用“运行模拟”对话框或“模拟构建器”，前者主要用于从模拟计划运行模拟。

要使用“运行模拟”对话框:

1. 从菜单中选择:

   分析 > 模拟...
2. 单击**打开现有模拟计划**。
3. 确保未选中**在模拟构建器中打开**复选框，并单击**继续**。
4. 打开模拟计划。
5. 在“运行模拟”对话框中单击**运行**。

要从“模拟构建器”运行模拟，请执行以下操作:

1. 从菜单中选择:

   分析 > 模拟...
2. 单击**打开现有模拟计划**。
3. 选中**在模拟构建器中打开**复选框，并单击**继续**。
4. 打开模拟计划。
5. 在“模拟”选项卡上修改所有需要修改的设置。
6. 单击**运行**以运行模拟。

此外，您还可以进行下列操作:

- 设置或修改敏感度分析，以调查改变固定输入值或模拟输入的分布参数所带来的影响。
- 将模拟输入分布和相关性重新拟合到新数据。
- 更改模拟输入的分布。
- 自定义输出。
- 将模拟数据保存为数据文件。

模拟构建器

“模拟构建器”提供了设计和运行模拟的完整功能集。它允许您执行以下常规任务:

- 针对 PMML 模型文件中定义的 IBM SPSS 模型，设计并运行模拟。
- 为由您指定的一组自定义方程所定义的预测模型设计并运行模拟。
- 设计并运行在没有预测模型的情况下生成数据的模拟。
- 基于现有模拟计划运行模拟，或者修改任何计划设置。

“模型”选项卡

对于基于预测模型的模拟，“模型”选项卡指定该模型的源。对于未包括预测模型的模拟，“模型”选项卡指定要模拟的字段。

选择 SPSS 模型文件。此选项指定在 IBM SPSS 模型文件中定义预测模型。IBM SPSS 模型文件是一个 XML 文件，其中包含根据 IBM SPSS Statistics 或 IBM SPSS Modeler 创建的模型 PMML。预测模型由过程创建，例如 IBM SPSS Statistics 中的线性回归和决策树，并且可以导出到模型文件中。您可以通过单击**浏览**并浏览到所需文件来使用其他模型文件。

模拟支持的 PMML 模型

- 线性回归
- 广义线性模型
- 一般线性模型
- 二元 Logistic 回归
- 多项 Logistic 回归
- 有序多项回归
- Cox 回归
- 树
- Boosted 树 (C5)
- 判别
- 两步聚类
- K-平均值聚类
- 类神经网络
- 规则集（决策列表）

注:

- 在模拟中不支持使用具有多个目标字段（变量）或拆分的 PMML 模型。
- 在模型中，二元 Logistic 回归模型的字符串输入值限长 8 个字节。如果要将这样的字符串输入与活动数据集拟合，请确保数据中值的长度不超过 8 个字节。长度超过 8 个字节的数据值将从该输入的相关分类分布中排除，并在“未匹配的类别”输出表中显示为“未匹配”。

输入模型方程。此选项指定预测模型由您创建的一个或多个自定义方程组成。通过单击**新建方程**可以创建方程。这将打开“方程编辑器”。您可以修改现有方程，复制它们以作为新方程的模板，对其重新排序和删除它们。

- “模拟构建器”不支持方程组系统或在目标变量中非线性的方程。
- 自定义方程将按其被指定的顺序进行评估。如果给定目标的方程依赖于另一个目标，那么该其他目标必须先由前一个方程来定义。

  例如，在下面的这三个方程中，*profit* 的方程依赖于 *revenue* 和 *expenses* 的值，因此 *revenue* 和 *expenses* 的方程必须位于 *profit* 的方程之前。

```
revenue = price*volume

expenses = fixed + volume*(unit_cost_materials + unit_cost_labor)

profit = revenue - expenses
```

在没有模型的情况下创建模拟数据。选择此选项可以在没有预测模型的情况下模拟数据。通过从活动数据集中选择字段，或者通过单击**新建**以定义新字段，指定要模拟的字段。

方程编辑器

“方程编辑器”允许您为预测模型创建或修改自定义方程。

- 方程表达式可以包含活动数据集中的字段，或者您在“方程编辑器”中定义的新输入字段。
- 您可以指定目标的属性，例如测量级别、值标签以及是否为目标生成输出。
- 您可以使用之前所定义方程的目标作为当前方程的输入，这将允许您创建耦合方程。
- 您可以为方程附加描述性注释。在“模型”选项卡上方程和注释将一起显示。

1. 输入目标名称。（可选）单击“目标”文本框下方的**编辑**以打开“定义的输入”对话框，这将允许您更改目标的缺省属性。
2. 要构建一个表达式，可以将成分粘贴到“数值表达式”字段中或者在“数值表达式”字段中直接输入。

- 您可以使用活动数据集中的字段来构建表达式，也可以通过单击**新建**按钮来定义新输入。这将打开“定义的输入”对话框。
- 通过从“函数组”列表中选择组，然后双击“函数”列表中的函数（或选择函数，然后单击“函数组”列表相邻的箭头），可以粘贴函数。请输入所有标注了问号的参数。标注了**全部**的函数组提供了所有可用函数的列表。对话框的保留区域中显示对当前所选函数的简要描述。
- 字符串常数必须包含在引号中。
- 如果值包含小数，那么必须使用句号 (.) 作为小数指示符。

注意: 模拟不支持带有字符串目标的定制方程。

定义的输入: “定义的输入”对话框允许您定义新输入并设置目标属性。

- 如果要在方程中使用的输入在活动数据集中不存在，那么必须先定义该输入，然后才能在方程中使用。
- 如果要在没有预测模型的情况下模拟数据，那么必须定义所有在活动数据集中不存在的模拟输入。

名称。指定目标或输入的名称。

目标。您可以指定目标的测量级别。缺省测量级别为连续。您还可以指定是否为此目标创建输出。例如，对于一组耦合方程，您可能只对最后一个方程的目标输出感兴趣，因此您可以不显示其他目标的输出。

要模拟的输入。此选项指定将按照在“模拟”选项卡上指定的概率分布来模拟输入值。测量级别确定一组缺省分布，在寻找与输入数据拟合最紧密的分布时（在“模拟”选项卡上单击**拟合**或**全部拟合**），将对这些缺省分布加以考虑。例如，如果测量级别为连续，那么将考虑正态分布（适用于连续数据），而不会考虑二项式分布。

注: 对于字符串输入，请选择“字符串”测量级别。要模拟的字符串输入限于“分类”分布。

固定值输入。这将指定输入值为已知，并且将固定为已知值。固定输入可为数值或字符串。为固定输入指定值。字符串值不应包含在引号中。

值标签。您可以为目标、模拟输入和固定输入指定值标签。值标签在输出图表和表中使用。

“模拟”选项卡

“模拟”选项卡指定除预测模型以外的所有模拟属性。在“模拟”选项卡上可以执行以下常规任务:

- 指定模拟输入的概率分布和固定输入值。
- 指定模拟输入之间的相关性。对于分类输入，可以指定为这些输入生成数据时，使用活动数据集中这些输入之间存在的关联。

- 指定高级选项，如尾部抽样和拟合分布到历史数据的标准。
- 自定义输出。
- 指定模拟计划的保存位置，并可选择保存模拟数据。

模拟字段

要运行模拟，必须将每个输入字段都指定为固定输入或模拟输入。模拟输入是其值不确定，并将从指定的概率分布抽取生成的这些值。存在所要模拟的输入的历史数据时，可以自动确定与该数据拟合最紧密的分布以及这些输入之间的任何相关性。如果历史数据不可用，或者您需要特殊的分布或相关性，您也可以手动指定分布或相关性。

固定输入是这些其值已知，并在模拟生成的每个个案中保持不变的值。例如，您拥有销售的线性回归模型作为输入数量的函数，包括价格，并且您希望将价格固定为当前市场价格。然后，您将价格指定为固定输入。

对于基于预测模型的模拟，模型中的每个预测变量都是该模拟的输入字段。对于未包括预测模型的模拟，“模型”选项卡上指定的字段是该模拟的输入。

自动拟合分布，并计算模拟输入的相关性。如果活动数据集包含您要模拟的输入的历史数据，那么可以自动找到与这些输入的数据拟合最紧密的分布，并确定它们之间的任何相关性。步骤如下所示:

1. 验证您要模拟的每个输入是否与活动数据集中的正确字段匹配。这些输入将在“输入”列中列出，“拟合到”列将显示活动数据集中的匹配字段。您可以通过从“拟合到”下拉列表中选择不同项目，将输入匹配到活动数据集中的不同字段。

   “拟合到”列中的 -无- 值表示输入无法与活动数据集中的字段自动匹配。缺省情况下，输入按名称、测量级别和类型（数字或字符串）与数据集字段匹配。如果活动数据集不包含输入的历史数据，那么需要手动指定输入分布或将输入指定为固定输入，如下所述。
2. 单击**全部拟合**。

拟合最紧密的分布及其相关参数显示在“分布”列中，并在历史数据的直方图（或条形图）上叠加显示分布图。模拟输入之间的相关性显示在“相关性”设置上。通过选择输入行并单击**拟合详细信息**，您可以检查拟合结果并定制特定输入的自动分布拟合。请参阅 第 156 页的『拟合详细信息』主题以获取更多信息。

通过选择输入行并单击**拟合**，可以针对特定输入运行自动分布拟合。所有与活动数据集中字段匹配的模拟输入之间的相关性也将自动计算。

注:

- 缺少任何模拟输入值的个案都会从分布拟合、相关性计算以及可选列联表计算（针对具有分类分布的输入）中排除。您可以选择指定是否将具有分类分布的输入的用户缺失值视为有效。缺省情况下，这些值视为缺失。有关更多信息，请参阅第 157 页的『高级选项』。
- 对于连续输入和有序输入，如果找不到任何所检验分布的可接受拟合，那么建议使用“经验”分布作为最紧密的拟合。对于连续输入，经验分布是历史数据的累积分布函数。对于有序输入，“经验”分布是历史数据的分类分布。

手动指定分布。您可以从**类型**下拉列表中选择分布，并在“参数”网格中输入分布参数，从而手动指定任何模拟输入的概率分布。一旦您输入分布参数后，将在“参数”网格旁边基于所指定的参数显示样本分布图。以下是有关特定分布的一些说明:

- **分类**。分类分布描述具有固定数目的值（称为类别）的输入字段。每个类别具有关联的概率，所有类别的概率总和等于 1。要输入类别，请单击“参数”网格中的左侧列，并指定类别值。在右侧列中输入该类别的关联概率。

注: 来自 PMML 模型的分类输入具有根据该模型确定的类别，这些类别不可修改。

- **负二项式 - 失败**。描述在试验序列中在观察到指定成功次数之前失败次数的分布。参数 *thresh* 是指定的成功次数，而参数 *prob* 是任何给定试验中的成功概率。
- **负二项式 - 试验**。描述在观察到指定成功次数之前所需试验次数的分布。参数 *thresh* 是指定的成功次数，而参数 *prob* 是任何给定试验中的成功概率。
- **范围**。该分布由一组区间组成，并为每个区间指定关联的概率，所有区间的概率总和等于 1。给定区间内的值抽取自在该区间上定义的均匀分布。通过输入最小值、最大值和关联概率即可指定区间。

  例如，您认为原材料成本有 40% 的可能性为每件产品 $10 - $15，有 60% 的可能性为每件产品 $15 - $20。那么，您可以使用范围分布来对成本建模，其中包括两个区间 [10 - 15] 和 [15 - 20]，并将第一个区间的关联概率设为 0.4，将第二个区间的关联概率设为 0.6。这些区间不需要是连续的，它们甚至可以重叠。例如，您可以指定区间为 $10 - $15 和 $20 - $25 或 $10 - $15 和 $13 - $16。
- **Weibull**。参数 *c* 是可选的位置参数，用于指定分布原点的所在位置。

下列分布的参数含义与“计算变量”对话框中相关联随机变量函数的参数含义相同：Bernoulli、Beta、二项式、指数、伽玛、对数正态、负二项式（试验和失败）、正态、泊松和均匀。

指定固定输入。通过从“分布”列中的**类型**下拉列表中选择“固定”并输入固定值，可以指定固定输入。值可以是数值或字符串，取决于输入是数值还是字符串。字符串值不应包含在引号中。

指定模拟值的边界。大多数分布支持指定模拟值的上限和下限。您可以在**最小值**文本框中输入值以指定下限，并可以在**最大值**文本框中输入值以指定上限。

锁定输入。通过选中带有锁图标的列中的复选框，可以锁定输入，从而将该输入排除在自动分布拟合之外。当您手动指定分布或固定值，并希望确保其不受自动分布拟合影响时，这非常有用。另外，如果您打算提供模拟计划给其他用户共享以供他们在“运行模拟”对话框中运行，并且希望禁止对特定输入进行任何更改，那么也最好进行锁定。在此情况下，无法在“运行模拟”对话框中修改对锁定输入的指定。

敏感度分析。敏感度分析允许您为每个指定值生成一组独立的模拟个案（即，另一模拟），从而调查固定输入中的或者模拟输入的分布参数中的系统性变化的效应。要指定敏感度分析，请选择固定输入或模拟输入，然后单击**敏感度分析**。敏感度分析局限于单一固定输入或者模拟输入的单一分布参数。请参阅第 156 页的『敏感度分析』主题以获取更多信息。

“拟合”状态图标

“拟合到”列中的图标表示每个输入字段的拟合状态。

表 3. 状态图标.

图标	描述
	尚未为输入指定分布，并且该输入未被指定为固定。要运行模拟，您必须为该输入指定分布，或将其定义为固定并指定固定值。
	该输入之前被拟合到活动数据集中不存在的字段。无需执行任何操作，除非您打算将输入的分布重新拟合到活动数据集。
	最紧密的拟合分布已替换为“拟合详细信息”对话框中的另一分布。

表 3. 状态图标 (续).

图标	描述
	输入已设置为最紧密的拟合分布。
	分布已手动指定或敏感度分析迭代已为此输入指定。

拟合详细信息: “拟合详细信息”对话框显示拟合特定输入的自动分布拟合结果。这些分布按拟合优度排序，最紧密的拟合分布列在最前面。通过在“使用”列中选择所需分布的相应单选按钮，您可以覆盖最紧密的拟合分布。选择“使用”列中的单选按钮还可显示在该输入的历史数据直方图（或条形图）上叠加的分布图。

拟合统计。缺省情况下，对于连续字段，使用 Anderson-Darling 检验来确定拟合优度。此外，仅对于连续字段，您还可以在“高级选项”设置中指定使用 Kolmogorov-Smirnoff 检验来确定拟合优度。对于连续输入，这两种检验的结果显示在“拟合统计”列中（A 为 Anderson-Darling; K 为 Kolmogorov-Smirnoff），并采用所选检验来对分布进行排序。对于有序和名义输入，使用卡方检验。与检验关联的 p 值也会显示。

参数。在“参数”列中显示了每个与拟合分布关联的分布参数。下列分布的参数含义与“计算变量”对话框中相关联随机变量函数的参数含义相同：Bernoulli、Beta、二项式、指数、伽玛、对数正态、负二项式（试验和失败）、正态、泊松和均匀。请参阅 主题以获取更多信息。对于分类分布，参数名称是类别，参数值是相关联的概率。

使用定制分布集重新拟合。缺省情况下，使用测量级别来确定自动分布拟合所考虑的分布集。例如，在拟合连续输入时，将考虑对数正态和伽玛之类的连续分布，而不会考虑泊松和二项式之类的离散分布。您可以在“重新拟合”列中选择分布，以便选择缺省分布的子集。您还可以从**视为（测量）**下拉列表中选择其他测量级别，并在“重新拟合”列中选择分布，从而覆盖缺省分布集。单击**运行重新拟合**以使用定制分布集进行重新拟合。

注:

- 缺少任何模拟输入值的个案都会从分布拟合、相关性计算以及可选列联表计算（针对具有分类分布的输入）中排除。您可以选择指定是否将具有分类分布的输入的用户缺失值视为有效。缺省情况下，这些值视为缺失。有关更多信息，请参阅第 157 页的『高级选项』。
- 对于连续输入和有序输入，如果找不到任何所检验分布的可接受拟合，那么建议使用“经验”分布作为最紧密的拟合。对于连续输入，经验分布是历史数据的累积分布函数。对于有序输入，“经验”分布是历史数据的分类分布。

敏感度分析: 敏感度分析允许您调查在指定值集范围内改变固定输入或模拟输入的分布参数所带来的影响。为每个指定值生成独立的模拟个案集，也即单独的模拟，这将允许您调查改变输入的影响。模拟个案的每个集合称为一次**迭代**。

迭代。此选项允许您指定输入变化的值集。

- 如果您要改变某个分布参数的值，请从下拉列表中选择参数。在“迭代参数值”网格中输入值集。单击**继续**会将指定的值添加到相关输入的“参数”网格中，并通过一个索引指定该值的迭代编号。
- 对于分类和范围分布，可以改变各个类别或区间的概率，但类别值和区间端点不得改变。请从下拉列表中选择类别或区间，并在“迭代参数值”网格中指定概率集。其他类别区间的概率将自动做出相应调整。

无迭代。使用此选项来取消输入迭代。单击**继续**将除去这些迭代。

相关性

要模拟的输入字段通常已确定相关，例如，身高与体重。为了确保模拟值保留这些相关性，必须考虑要模拟的输入之间的相关性。

拟合时重新计算相关性。此选项指定，通过“模拟字段”设置中的**全部拟合**或**拟合**操作将分布与活动数据集拟合时，将自动计算模拟输入之间的相关性。

拟合时不重新计算相关性。如果您打算手动指定相关性，并且禁止在自动将分布拟合到活动数据集时覆盖相关性，请选择此项。在“相关性”网格中输入的值必须介于 -1 与 1 之间。值为 0 指定在相关输入对之间不存在相关性。

重置。这会将所有相关性重置为 0。

对具有分类分布的输入使用拟合的多向列联表。对于具有分类分布的输入，可以根据描述了这些输入之间的关联的活动数据集自动计算多向列联表。然后，在为这些输入生成数据时，使用这个列联表。如果您选择保存模拟计划，那么这个列联表将保存在计划文件中，并在您运行该计划时使用。

- **根据活动数据集计算列联表**。如果您正在使用其中包含列联表的现有模拟计划，那么可以根据活动数据集重新计算该列联表。此操作将覆盖已装入的计划文件中的列联表。
- **使用装入的模拟计划中的列联表**。缺省情况下，在装入其中包含列联表的模拟计划时，将使用该计划中的列联表。通过选择**根据活动数据集计算列联表**，可以根据活动数据集重新计算列联表。

高级选项

个案的最大数量。此选项指定要生成的模拟数据及关联目标值的最大个案数目。如果指定了敏感度分析，此为每次迭代的最大个案数目。

停止标准的目标。如果预测模型包含多个目标，那么可以选择要对其应用停止标准的目标。

停止标准。这些选项指定在生成最大允许个案数目之前可能停止模拟的标准。

- **持续至达到最大数量**。此选项指定将持续生成模拟个案，直至达到最大个案数目。
- **对尾部取样后停止**。如果您希望确保对指定目标分布的一个尾部充分抽样，请使用此选项。这将持续生成模拟个案，直至完成指定的尾部抽样或达到最大个案数目。如果预测模型包含多个目标，请从**停止标准的目标**下拉列表中选择要对其应用此标准的目标。

  类型。您可以通过指定目标值（如 10,000,000）或百分位数（如第 99 个百分位数）来定义尾部区域的边界。如果您在**类型**下拉列表中选择“值”，请在“值”文本框中输入边界值，并使用**左/右侧**下拉列表来指定这是左侧还是右侧尾部区域的边界。如果在**类型**下拉列表中选择了“百分位数”，请在“百分位数”文本框中输入值。

  频率。指定必须处于尾部区域内的目标值数量，以确保对尾部充分抽样。当达到该数量时，将停止生成个案。
- **当平均值置信区间处于指定阈值内时停止**。如果您希望确保给定目标的平均值具有指定精确度，请使用此选项。这将持续生成模拟个案，直至获得指定的精确度或达到最大个案数目。要使用此选项，应指定置信度和阈值。模拟个案将持续生成，直至指定水平的关联置信区间处于阈值内。例如，您可以使用此选项来指定持续生成个案，直至 95% 置信度的平均值置信区间处于平均值的 5% 以内。如果预测模型包含多个目标，请从**停止标准的目标**下拉列表中选择要对其应用此标准的目标。

  阈值类型。您可以将阈值指定为数值或平均值百分比。如果在**阈值类型**下拉列表中选择了“值”，请在“阈值（值）”文本框中输入阈值。如果在**阈值类型**下拉列表中选择了“百分比”，请在“阈值（百分比）”文本框中输入值。

要取样的个案数。此选项指定在将模拟输入分布拟合到活动数据集时要使用的个案数。如果您的数据集非常大，您可能会考虑限制要用于分布拟合的个案数。如果您选择**限制为 N 个个案**，那么将使用前 N 个个案。

拟合优度标准（连续）。对于连续输入，您可以在将模拟输入分布拟合到活动数据集时，使用 Anderson-Darling 或 Kolmogorov-Smirnoff 拟合优度检验来排序分布。缺省选择 Anderson-Darling 检验，并且如果您希望确保在尾部区域的最佳拟合，建议您使用该检验。

经验分布。对于连续输入，经验分布是历史数据的累积分布函数。您可以指定要用于计算连续输入的经验分布的分箱数量。缺省值为 100，最大值为 1000。

重复结果。设置随机种子允许您复制模拟。指定一个整数，或单击**生成**，这将产生一个介于 1 与 2147483647 之间（包括 1 和 2147483647）的伪随机整数。缺省值为 629111597。

具有分类分布的输入的用户缺失值。这些控件指定是否将具有分类分布的输入的用户缺失值视为有效。所有其他类型的输入的系统缺失值和用户缺失值始终被视为无效。所有输入都必须具有有效的值，这样才能将个案包括在分布拟合、相关性计算以及可选列联表的计算中。

密度函数

这些设置允许您为连续目标自定义概率密度函数和累积分布函数的输出，以及为分类目标自定义预测值的条形图。

概率密度函数 (PDF)。概率密度函数显示目标值的分布。对于连续目标，它允许您确定目标处于给定区域内的概率。对于分类目标（具有名义或有序测量级别的目标），将生成条形图以显示处于每个目标类别内的个案百分比。还提供了适合 PMML 模型分类目标的其他选项，并带有以下所述的“要报告的类别值”设置。

对于两步聚类模型和 K-Means 聚类模型，将生成聚类成员条形图。

累积分布函数 (CDF)。累积分布函数显示目标值小于或等于指定值的概率。它仅对连续目标可用。

滑块位置。您可以指定 PDF 和 CDF 图表中可移动的参考线的初始位置。对下限线条和上限线条指定的值是指水平轴上的位置，而不是百分位数。您可以通过选择**负无穷**除去下方的线条，也可以通过选择**无穷**除去上方的线条。缺省情况下，这些线条定位在第 5 个和第 95 个百分位数位置。如果由于敏感度分析迭代产生了多个目标或结果而在单个图表上显示多个分布函数，那么缺省分布是指第一次迭代或第一个目标的分布。

参考线（连续）。您可以要求为连续目标的概率密度函数和累积分布函数添加各种垂直参考线。

- **西格玛**。您可以在相对于目标平均值加/减指定标准差数的位置添加参考线。
- **百分位数**。通过在“底部”和“顶部”文本框中输入值，可以在目标分布的一个或两个百分位值处添加参考线。例如，“顶部”文本框中的值 95 表示第 95 个百分位数，即，95% 的观察值位于此值之下。同样，“底部”文本框中的值 5 表示第五个百分位数，即，5% 的观察值位于此值之下。
- **定制参考线**。您可以在指定目标值处添加参考线。

注: 如果由于敏感度分析迭代产生了多个目标或结果而在单个图表上显示多个分布函数，那么参考线仅应用于第一次迭代或第一个目标的分布。您可以使用“图表选项”对话框（通过 PDF 或 CDF 图表进行访问）为其他分布添加参考线。

重叠不同连续目标的结果。在多个连续目标的情况下，此选项指定是否在单个图表上显示所有此类目标的分布函数，其中概率密度函数和累积分布函数分别对应一个图表。如果不选择此选项，每个目标的结果将显示在单独的图表上。

要报告的类别值。对于具有分类目标的 PMML 模型，模型结果是一组有关目标值处于每个类别中的预测概率，每个类别一个概率。具有最高概率的类别将作为预测类别，并用于生成前面的**概率密度函数**设置中描述的条形图。选择**预测类别**将生成该条形图。选择**预测概率**将为目标的每个类别生成预测概率的分布直方图。

针对敏感度分析进行分组。包含敏感度分析的模拟将为该分析定义的每次迭代（当前改变输入的每个值对应一次迭代）生成独立的预测目标值集。如果存在迭代，分类目标的预测类别条形图将显示为包含所有迭代结果的复式条形图。您可以选择对类别或迭代进行分组。

输出

龙卷风图表。龙卷风图表以条形图形式通过多种度量来显示目标和模拟输入之间的关系。

- **目标与输入的相关性**。此选项创建给定目标与其每个模拟输入之间相关系数的龙卷风图表。此类龙卷风图表不支持具有名义或有序测量级别的目标或者具有分类分布的模拟输入。
- **对方差的贡献**。此选项创建龙卷风图表以显示每个模拟输入对目标方差的贡献，这将允许您评估每个输入对目标总体不确定性的贡献程度。此类龙卷风图表不支持具有有序或名义测量级别的目标或者具有下列任何分布的模拟输入：分类、Bernoulli、二项式、泊松或负二项式。
- **目标对变化的敏感度**。此选项创建龙卷风图表以通过加上或减去指定数量的输入关联分布的标准差来显示调整每个模拟输入对目标的影响。此类龙卷风图表不支持具有有序或名义测量级别的目标或者具有下列任何分布的模拟输入：分类、Bernoulli、二项式、泊松或负二项式。

目标分布箱图。箱图对连续目标可用。如果预测模型具有多个连续目标，并且您希望在单个图表上显示所有目标的箱图，请选中**重叠不同目标的结果**。

目标与输入散点图。目标与模拟输入散点图对连续和分类目标均可用，其中包括具有连续和分类输入的目标散点。涉及分类目标或分类输入的散点显示为热图。

创建百分位值表。对于连续目标，您可以获得目标分布的指定百分位数表。四分位数（第 25、50、75 个百分位数）将观察值分为四个大小相等的组。如果您希望相等组的数目不等于 4，请选择**区间**并指定数目。选择**定制百分位数**以指定各个百分位数，例如第 99 个百分位数。

目标分布的描述统计。此选项创建连续和分类目标以及连续输入的描述统计表。对于连续目标，此表包括平均值、标准差、中位数、最小值和最大值、指定水平的平均值置信区间，以及目标分布的第 5 和 95 个百分位数。对于分类目标，此表包括处于每个目标类别中的个案百分比。对于 PMML 模型的分类目标，此表还包括每个目标类别的平均值概率。对于连续输入，此表包括平均值、标准差、最小值和最大值。

输入的相关性和列联表。此选项显示模拟输入之间的相关系数的表。根据列联表生成具有分类分布的输入时，还将显示为这些输入生成的数据的列联表。

要包括在输出中的模拟输入。缺省情况下，输出中包含所有模拟输入。您可以从输出中排除选定模拟输入。这将从龙卷风图表、散点图和表格输出中排除这些模拟输入。

连续目标的限制范围。您可以为一个或多个连续目标指定有效值的范围。处于指定范围之外的值将从所有与目标相关联的输出和分析中排除。要设置下限，请在“限制”列中选择**下限**，并在“最小值”列中输入值。要设置上限，请在“限制”列中选择**上限**，并在“最大值”列中输入值。要同时设置下限和上限，请在“限制”列中选择**两者**，并在“最小值”和“最大值”列中输入值。

显示格式。您可以设置目标和输入值（固定输入和模拟输入）的显示格式。

保存

保存此模拟的计划。您可以将当前的模拟规范保存为模拟计划文件。模拟计划文件具有 *.splan* 扩展名。您可以在“模拟构建器”中重新打开该计划，并可选择做出修改，然后运行模拟。您可以将模拟计划共享给其他用户，

然后他们可以在“运行模拟”对话框中加以运行。模拟计划包含除下列各项以外的所有规范：密度函数的设置、图表和表的输出设置、拟合的高级选项设置、经验分布以及随机种子。

将模拟数据保存为新数据文件。您可以将模拟输入、固定输入和预测目标值保存到 SPSS Statistics 数据文件、当前会话中的新数据集或者 Excel 文件。数据文件的每个个案（或行）都由目标的预测值以及生成目标值的模拟输入和固定输入组成。指定敏感度分析后，每次迭代都将生成标注了迭代编号的个案的连续集。

“运行模拟”对话框

“运行模拟”对话框适合那些拥有模拟计划且主要打算运行模拟的用户。它还提供了一些功能以允许您在不同条件下运行模拟。它允许您执行以下常规任务：

- 设置或修改敏感度分析，以调查改变固定输入值或模拟输入的分布参数所带来的影响。
- 将不确定输入的概率分布（以及这些输入之间的相关性）重新拟合到新数据。
- 修改模拟输入的分布。
- 自定义输出。
- 运行模拟。

“模拟”选项卡

“模拟”选项卡允许您指定敏感度分析、将模拟输入的概率分布与这些输入之间的相关性重新拟合到新数据，以及修改与模拟输入关联的概率分布。

对于模拟计划中定义的每个输入字段，“模拟输入”网格都包含一个条目。每个条目显示输入名称及其关联的概率分布类型，以及关联分布曲线的样本图。每个输入还具有关联的状态图标（带有勾选标记的彩色圆圈），这在您重新将分布拟合到新数据时非常有用。此外，输入还可能包含一个锁图标，它表示该输入在“运行模拟”对话框中被锁定，无法修改或重新拟合到新数据。要修改已锁定的输入，您需要在“模拟构建器”中打开模拟计划。

每个输入可以是模拟或固定输入。模拟输入是其值不确定，并将从指定的概率分布抽取生成的这些值。固定输入是这些其值已知，并在模拟生成的每个个案中保持不变的值。要使用特定输入，请在“模拟输入”网格中选择相应的条目。

指定敏感度分析

敏感度分析允许您为每个指定值生成一组独立的模拟个案（即，另一模拟），从而调查固定输入中的或者模拟输入的分布参数中的系统性变化的效应。要指定敏感度分析，请选择固定输入或模拟输入，然后单击**敏感度分析**。敏感度分析局限于单一固定输入或者模拟输入的单一分布参数。请参阅第 156 页的『敏感度分析』主题以获取更多信息。

将分布重新拟合到新数据

要将模拟输入的概率分布（以及这些输入之间的相关性）自动重新拟合到新数据：

1. 验证每个模型输入是否与活动数据集中的正确字段匹配。每个模拟输入都与活动数据集中的特定字段拟合，该字段在该输入的相关**字段**下拉列表中指定。您可以寻找带有勾选标记及问号的状态图标以方便确定不匹配的输入，如下所示。

2. 通过选择**与数据集中的字段拟合**并从列表中选择字段，可以修改任何必要的字段匹配。
3. 单击**全部拟合**。

对于拟合的每个输入，在显示与数据拟合最紧密的分布时，还将显示叠加在历史数据直方图（或条形图）上的分布图。如果找不到可接受的拟合，那么会使用经验分布。对于拟合到经验分布的输入，您将只能看到历史数据的直方图，因为经验分布实际上是由该直方图表示的。

注意: 有关状态图标的完整列表，请参阅第 154 页的『模拟字段』主题。

修改概率分布

您可以修改模拟输入的概率分布，并可选择将模拟输入更改为固定输入，反之亦可。

1. 选择输入，并选择**手动设置分布**。
2. 选择分布类型，并指定分布参数。要将模拟输入更改为固定输入，请在**类型**下拉列表中选择“固定”。

一旦您输入分布参数后，在输入条目中显示的样本分布图将更新以反映您的更改。有关手动指定概率分布的更多信息，请参阅第 154 页的『模拟字段』主题。

拟合时包括分类输入的用户缺失值。此选项指定与活动数据集中的数据进行重新拟合时，是否将具有分类分布的输入的用户缺失值视为有效。所有其他类型的输入的系统缺失值和用户缺失值始终被视为无效。所有输入都必须具有有效的值，这样才能将个案包括在分布拟合以及相关性计算中。

“输出”选项卡

“输出”选项卡允许您自定义由模拟生成的输出。

密度函数。密度函数是探测模拟结果集的主要手段。

- **概率密度函数**。概率密度函数显示目标值的分布，这将允许您确定目标处于给定区域内的概率。对于具有固定成果集（例如，“服务较差”、“服务尚可”、“服务良好”和“服务最佳”）的目标，将生成条形图，以显示处于每个目标类别中的个案百分比。
- **累积分布函数**。累积分布函数显示目标值小于或等于指定值的概率。

龙卷风图表。龙卷风图表以条形图形式通过多种度量来显示目标和模拟输入之间的关系。

- **目标与输入的相关性**。此选项创建给定目标与其每个模拟输入之间相关系数的龙卷风图表。
- **对方差的贡献**。此选项创建龙卷风图表以显示每个模拟输入对目标方差的贡献，这将允许您评估每个输入对目标总体不确定性的贡献程度。
- **目标对变化的敏感度**。此选项创建龙卷风图表以通过加上或减去输入关联分布的一个标准差来显示调整每个模拟输入对目标的影响。

目标与输入散点图。此选项生成目标与模拟输入的散点图。

目标分布箱图。此选项生成目标分布的箱图。

四分位数表。此选项生成目标分布的四分位数表。分布的四分位数包括分布的第 25、50、75 个百分位数，它们将观察值分为四个大小相等的组。

输入的相关性和列联表。此选项显示模拟输入之间的相关系数的表。如果模拟计划指定了根据列联表生成分类数据，那么将显示具有分类分布的输入之间的关联的列联表。

重叠不同目标的结果。如果当前模拟的预测模型包含多个目标，您可以指定是否在单个图表上显示来自不同目标的结果。该设置适用于概率密度函数、累积分布函数图表和箱图。例如，如果您选择此项，那么会在单个图表上显示所有目标的概率密度函数。

保存此模拟的计划。您可以将任何模拟修改保存到模拟计划文件中。模拟计划文件具有 *.splan* 扩展名。您可以在“运行模拟”对话框或“模拟构建器”中重新打开该计划。模拟计划包含除输出设置外的所有规范:

将模拟数据保存为新数据文件。您可以将模拟输入、固定输入和预测目标值保存到 SPSS Statistics 数据文件、当前会话中的新数据集或者 Excel 文件。数据文件的每个个案（或行）都由目标的预测值以及生成目标值的模拟输入和固定输入组成。指定敏感度分析后，每次迭代都将生成标注了迭代编号的个案的连续集。

如果您需要自定义此处未提及的其他输出属性，可以考虑在“模拟构建器”中运行模拟。请参阅 第 151 页的『从模拟计划运行模拟』主题以获取更多信息。

使用模拟图表输出

模拟所生成的许多图表具有交互式功能，允许您自定义显示内容。在“输出查看器”中激活（双击）图表对象，即可使用交互式功能。所有模拟图表均为图形画板可视化。

连续目标的概率密度函数图表。该图表具有两个可滑动的垂直参考线，将图表划分成不同区域。图表下面的表格显示目标处于每个区域中的概率。如果在同一图表上显示多个密度函数，那么表中设有单独行来显示每个密度函数的概率。每条参考线均设有滑块（倒三角形）以允许您方便地移动参考线。通过单击图表上的**图表选项**按钮，还可以使用许多其他功能。具体来说，您可以明确设置滑块位置、添加固定参考线，以及将图表视图从连续曲线更改为直方图（反之亦可）。请参阅 『图表选项』主题以获取更多信息。

连续目标的累积分布函数图表。与前面的概率密度函数图表相同，该图表也具有两个可移动的垂直参考线以及关联的表格。您也可以访问“图表选项”对话框，并从中明确设置滑块位置、添加固定参考线，并指定累积分布函数显示为递增函数（缺省）还是递减函数。请参阅 『图表选项』主题以获取更多信息。

具有敏感度分析迭代的分类目标条形图。对于具有敏感度分析迭代的分类目标，预测目标类别的结果显示为复式条形图，其中包括所有迭代的结果。该图表包含一个下拉列表，允许您按类别或迭代来聚类。对于两步聚类模型和 K-Means 聚类模型，您可以选择按聚类编号或迭代来聚类。

具有敏感度分析迭代的多目标箱图。对于具有多个连续目标和敏感度分析迭代的预测模型，选择在单个图表上显示所有目标的箱图将生成聚类箱图。该图表包含一个下拉列表，允许您按目标或迭代来聚类。

图表选项

“图表选项”对话框允许您自定义模拟所生成的概率密度函数和累积分布函数已激活图表的显示内容。

视图。**视图**下拉列表仅适用于概率密度函数图表。它允许您将图表视图从连续曲线切换到直方图。如果在同一图表上显示多个密度函数，该功能不可用。在此情况下，密度函数只能显示为连续曲线。

排序。**排序**下拉列表仅适用于累积分布函数图表。它指定累积分布函数显示为递增函数（缺省）还是递减函数。当显示为递减函数时，在水平轴给定点处的函数值为目标处于该点右侧的概率。

滑块位置。您可以在“上限”和“下限”文本框内输入值，以明确设置可滑动参考线的位置。您可以选择**负无穷**，即将位置设置为负无穷远，以除去左侧参考线；还可以选择**无穷**，即将位置设置为无穷远，以除去右侧参考线。

参考线。您可以对概率密度函数和累积分布函数添加各种固定的垂直参考线。如果由于敏感度分析迭代产生了多个目标或结果而在单个图表上显示多个函数，那么您可以指定这些参考线所应用于的特定函数。

- **西格玛**。您可以在相对于目标平均值加/减指定标准差数的位置添加参考线。
- **百分位数**。通过在“底部”和“顶部”文本框中输入值，可以在目标分布的一个或两个百分位值处添加参考线。例如，“顶部”文本框中的值 95 表示第 95 个百分位数，即，95% 的观察值位于此值之下。同样，“底部”文本框中的值 5 表示第五个百分位数，即，5% 的观察值位于此值之下。
- **定制位置**。您可以在沿水平轴的指定值处添加参考线。

标注参考线。此选项控制是否将标签应用于选择的参考线。

通过在“图表选项”对话框中取消选中相关选项，并单击**继续**，可以除去参考线。

第 35 章 地理空间建模

地理空间建模方法设计为发现数据中包含地理空间（地图）组件的模式。“地理空间建模向导”提供了用于分析带与不带时间组件的地理空间数据的方法。

基于事件和地理空间数据查找关联（地理空间关联规则）

通过使用地理空间关联规则，您可以基于空间和非空间属性在数据中查找模式。例如，您可以通过位置和人口统计信息属性在犯罪数据中识别模式。通过这些模式，您可以构建规则来预测可能发生某些类型犯罪的地点。

使用时间序列和地理空间数据进行预测（空间时间预测）

空间时间预测使用包含位置数据、预测（预测变量）输入字段、一个或多个时间字段和一个目标字段的数据。每个位置在数据中都有很多行，用于表示每个时间间隔每个预测变量和目标的值。

使用“地理空间建模向导”

1. 从菜单中选择:

   分析 > 空间和时间建模 > 空间建模

2. 执行向导中的步骤。

选择地图

地理空间建模可以使用一个或多个地图数据源。地图数据源包含用于定义地理区域和其他地理功能（例如公路或河流）的信息。许多地图源还包含人口统计信息或其他描述性数据与事件数据，例如犯罪报告或失业率。您可以使用先前定义的地图规范文件，也可以在此处定义地图规范并保存这些规范以供后续使用。

装入地图规范

装入先前定义的地图规范文件 (.mplan)。您在这里定义的地图数据源可以保存到地图规范文件中。对于空间时间预测，如果选择标识了多个地图的地图规范文件，将提示您从该文件中选择一个地图。

添加地图文件

添加 ESRI 形状文件 (.shp) 或包含 ESRI 形状文件的 .zip 归档。

- 在 .shp 文件所在的位置中必须有对应的 .dbf 文件，并且此文件必须具有与 .shp 文件相同的根名。
- 如果文件是 .zip 归档，.shp 和 .dbf 文件必须具有与 .zip 归档相同的根名。
- 如果没有对应的投影文件 (.prj)，将提示您选择投影系统。

关系 对于地理空间关联规则，此列用于定义事件如何与地图中的功能相关联。此设置不可用于空间时间预测。

上移/下移

地图元素的层顺序由它们在列表中显示的顺序决定。列表中的第一个地图是最低层。

选择地图

对于空间时间预测，如果选择标识了多个地图的地图规范文件，将提示您从该文件中选择一个地图。空间时间预测不支持多个地图。

© Copyright IBM Corporation 1989, 2014

地理空间关系

对于地理空间关联规则，“地理空间关系”对话框用于定义事件如何与地图中的功能相关联。

- 此设置仅应用于地理空间关联规则。
- 此设置仅影响与选择数据源的步骤上的上下文数据的地图相关联的数据源。

关系

接近 事件在地图上接近指定点或区域的位置处发生。

内部 事件在地图上指定区域内发生。

包含 事件区域包含地图上下文对象。

交叉 不同地图的线条或区域相互交叉的位置。

交汇 对于多个地图，指的是不同地图的线条（公路、河流、铁路）交汇的位置。

东南西北
事件发生于地图上指定点的北部、南部、东部或西部。

设置坐标系统

如果地图没有投影文件（.prj）或者您将数据源中的两个字段定义为一组坐标，那么必须设置坐标系统。

缺省地理（纵坐标和横坐标）
坐标系统为纵坐标和横坐标。

简单直角坐标（X 和 Y）
坐标系统为简单的 X 和 Y 坐标。

使用知名标识 (WKID)
“知名标识”用于常用投影。

使用坐标系统名称
坐标系统基于命名投影。名称括在括号中。

设置投影

如果无法通过地图随附的信息确定投影系统，那么需要指定投影系统。此条件的最常见原因是存在与无法使用的地图或投影文件相关联的投影（.prj）文件。

- **城市、区域或国家或地区 (Mercator)**
- **大的国家或地区、多个国家或地区或大陆 (Winkel Tripel)**
- **非常接近赤道的区域 (Mercator)**
- **接近极地的区域 (Stereographic)**

Mercator 投影是许多地图中的常用投影。此投影将地球视为在平面上铺开的柱面。Mercator 投影会误报大对象的大小和形状。此误报会随据赤道距离的增大以及极地距离的接近而增加。Winkel Tripel 和 Stereographic 投影对地图在二维中表示三维球体的比例因子进行调整。

投影和坐标系统

如果选择多个地图，并且这些地图有不同的投影和坐标系统，那么必须选择具有要使用的投影系统的地图。如果所有地图一起汇总到输出中，那么该投影系统将用于所有地图。

数据源

数据源可以是与形状文件一起提供的 dBase 文件、IBM SPSS Statistics 数据文件或者当前会话中打开的数据集。

上下文数据。上下文数据标识地图上的功能。上下文数据还可包含用作模型的输入的字段。要使用与地图形状文件 (.shp) 关联的上下文 dBase (.dbf) 文件，上下文 dBase 文件必须与该形状文件位于同一位置，并且具有相同的根名。例如，如果形状文件是 `geodata.shp`，那么 dBase 文件必须命名为 `geodata.dbf`

事件数据。事件数据包含与所发生的事件（例如，犯罪或事故）有关的信息。此选项仅可用于地理空间关联规则。

点密度。核心密度估算的时间间隔和坐标数据。此选项仅可用于地理空间时间预测。

添加。打开用于添加数据源的对话框。数据源可以是与形状文件一起提供的 dBase 文件、IBM SPSS Statistics 数据文件或者当前会话中打开的数据集。

关联。打开一个对话框以供指定用于将数据与地图关联的标识（坐标或键）。每个数据源必须包含一个或多个用于将数据与地图关联的标识。形状文件随附的 dBase 文件通常包含自动用作缺省标识的字段。对于其他数据源，您必须制定用作标识的字段。

验证键。打开用于验证地图与数据源之间的键匹配的对话框。

地理空间关联规则

- 至少有一个数据源必须包含事件数据源。
- 所有事件数据源都必须使用相同格式的地图关联标识：坐标或键值。
- 如果事件数据源与具有键值的地图关联，那么所有事件源都必须使用相同的地图功能类型（例如，多边形、点、线）。

空间时间预测

- 必须有一个上下文数据源。
- 如果只有一个数据源（无关联地图的数据文件），那么它必须包含坐标值。
- 如果您有两个数据源，那么一个数据源必须是上下文数据，另一个数据源必须是点密度数据。
- 不能包含两个以上的数据源。

添加数据源

数据源可以是与形状文件和上下文文件一起提供的 dBase 文件、IBM SPSS Statistics 数据文件或者当前会话中打开的数据集。

对于同一数据源，如果要使用其不同的空间关联，可以多次添加该数据源。

数据和地图关联

每个数据源必须包含一个或多个用于将数据与地图关联的标识。

坐标 数据源包含用于表示笛卡尔坐标的字段，选择用于表示 X 和 Y 坐标的字段。对于地理空间关联规则，还可以存在 Z 坐标。

键值 数据源中字段的键值对应于所选地图键。例如，区域地图可能在每个区域上都标注了名称标识（地图键）。该标识对应于数据中也包含区域名称的字段（数据键）。字段与地图键基于两个列表中的显示顺序进行匹配。

验证键

“验证键”对话框基于所选标识键提供地图与数据源之间的记录匹配摘要。如果有不匹配的数据键值，您可以手动将它们与键值匹配。

地理空间关联规则

对于地理空间关联规则，在定义地图和数据源之后，向导中的其余步骤包括:

- 如果有多个事件数据源，请定义事件数据源的合并方式。
- 选择用作分析中的条件和预测的字段。

根据需要，您可以:

- 选择不同输出选项。
- 保存评分模型文件。
- 为模型所用数据源中的预测值和规则创建新字段。
- 定制用于构建关联规则的设置。
- 定制分箱和聚集设置。

定义事件数据字段

对于地理空间关联规则，如果有多个事件数据源，将合并事件数据源。

- 缺省情况下，仅包含所有事件数据源通用的字段。
- 您可以显示通用字段列表、特定数据源字段或所有数据源字段，并选择您要包含的字段。
- 对于通用字段，**类型**和**测量**对于所有数据源都必须相同。如果存在冲突，您可以指定要用于每个通用字段的类型和测量级别。

选择字段

可用字段列表包含来自事件数据源的字段和来自上下文数据源的字段。

- 您可以通过从**数据源**列表选择数据源来控制所显示的字段列表。
- 您必须至少选择两个字段。至少一个条件字段，至少一个预测字段。可以通过多种方法来满足这一要求，包括为**两者（条件和预测）**列表选择两个字段。
- 关联规则预测基于条件字段值的预测字段的值。例如，在规则“If x=1 and y=2, then z=3”中，x 和 y 的值是条件，z 的值是预测。

输出

规则表 每个规则表显示置信度、规则支持、提升度、条件支持和可部署能力的顶级规则和值。每个表按照所选标准的值进行排序。您可以显示所有规则，也可以根据所选标准显示前（**数字**）条规则。

可排序字云

基于所选标准值的顶级规则列表。文本大小指示规则的相对重要性。交互式输出对象包含置信度、规则支持、提升度、条件支持和部署能力的顶级规则。所选标准确定缺省情况下显示哪列规则。您可以在输出中以交互方式选择不同的标准。**要显示的最大规则数**确定输出中显示的规则数。

地图 顶级规则的交互式条形图和地图，基于所选标准。每个交互式输出对象包含置信度、规则支持、提升度、条件支持和部署能力的顶级规则。所选标准确定缺省情况下显示哪列规则。您可以在输出中以交互方式选择不同的标准。**要显示的最大规则数**确定输出中显示的规则数。

模型信息表

字段转换

描述应用于分析中所使用的字段的转换。

记录摘要。

包含和排除的记录的数量与百分比。

规则统计信息。

条件支持、置信度、规则支持、提升度和部署能力的汇总统计。该统计信息包括平均值、最小值、最大值和标准差。

最频繁的项。

出现最频繁的项。项包含在规则的条件或预测中。例如，age < 18 或 gender=female。

最频繁的字段。

规则中出现最频繁的字段。

排除的输入。

从分析排除的字段以及每个字段的排除原因。

规则表、字云和地图的标准

置信度。

正确规则预测数的百分比。

规则支持。

规则为 true 的个案的百分比。例如，如果规则为“If x=1 and y=2, then z=3”，那么规则支持是指 x=1、y=2 并且 z=3 的个案在数据中所占的实际百分比。

提升度。

提升度用于测量与随机概率相比，规则改进预测的程度。这是正确预测数占所有预测值出现次数的比率。该值必须大于 1。例如，如果预测的值出现时间占了 20%，预测置信度为 80%，那么提升度值是 4。

条件支持。

规则条件存在的个案的百分比。例如，如果规则是“If x=1 and y=2, then z=3”，那么条件支持是指 x=1 并且 y=2 的个案在数据中所占的比例。

部署能力。

满足条件时不正确预测的比例。部署能力等于（1 - 置信度）乘以条件支持或条件支持减去规则支持。

保存

将地图和上下文数据保存为地图规范

将地图规范保存为外部文件 (.mplan)。您可以将此地图规范文件上载到向导以供后续分析。您还可以将地图规范文件与 SPATIAL ASSOCIATION RULES 命令一起使用。

将任何地图和数据文件复制到规范

地图规范中所用的地图形状文件、外部数据文件和数据集中的数据保存在地图规范文件中。

评分 将最佳规则值、规则置信度值、规则数字标识值保存为指定数据源中的新字段。

要评分的数据源

在其中创建新字段的一个或多个数据源。如果在当前会话中没有打开数据源，那么将在当前会话中将其打开。您必须明确保存修改后的文件以保存新字段。

目标值 为所选目标（预测）字段创建新字段。

- 为每个目标字段创建两个新字段：预测值和置信度值。

- 对于连续（刻度）目标字段，预测值是用于描述值范围的字符串。格式为“(value1, value2]”的值表示“大于 value1 且小于等于 value2”。

最佳规则数

为所指定的最佳规则数创建新字段。为每个规则创建三个新字段：规则值、置信度值和规则数字标识值。

名称前缀

要用于新字段名称的前缀。

规则构建

规则构建参数用于为所生成的关联规则生成标准。

每条规则的项数

规则条件和预测中可包含的字段值数量。总项数不能超过 10。例如，在规则“If x=1 and y=2, then z=3”中，有两个条件项和一个预测项。

最大预测数。

对于每个规则而言，预测中可以出现的最大字段值数量。

最大条件数。

对于每个规则而言，条件中可以出现的最大字段值数量。

排除对 排除指定字段对以不包含在同一规则中。

规则标准

置信度。

规则要包含在输出中而必须具有的最小置信度。置信度是正确预测数的百分比。

规则支持。

规则要包含在输出中而必须具有的最小规则支持。该值表示在所观察到的数据中规则为 true 的个案的百分比。例如，如果规则为“If x=1 and y=2, then z=3”，那么规则支持是指 x=1、y=2 并且 z=3 的个案在数据中所占的实际百分比。

条件支持。

规则要包含在输出中而必须具有的最小条件支持。该值表示条件存在的个案的百分比。例如，如果规则是“If x=1 and y=2, then z=3”，那么条件支持是指 x=1 并且 y=2 的个案在数据中所占的百分比。

提升度。

规则要包含在输出中而必须具有的最小提升度。提升度用于测量与随机概率相比，规则改进预测的程度。这是正确预测数占所有预测值出现次数的比率。例如，如果预测的值出现时间占了 20%，预测置信度为 80%，那么提升度值是 4。

视为相同

标识应视为相同字段的字段对。

分箱和聚集

- 当数据中的记录条数超过地图中的功能数时，需要聚集。例如，您具有各个县的数据记录，但您的地图是省/自治区/直辖市的地图。
- 您可以为连续和有序字段指定聚集汇总测量方法。名义字段基于最常见的值进行聚集。

连续 对于连续（刻度）字段，汇总测量可以是平均值、中间值或总和。

有序 对于有序字段，汇总测量可以是中间值、众数、最高值或最低值。

分箱数 为连续（刻度）字段设置最大分箱数。连续字段始终分组或“分箱”为值范围。例如: 小于等于 5、大于 5 且小于等于 10 或者大于 10。

聚集地图

对数据和地图应用聚集。

定制特定字段的设置

您可以覆盖特定字段的缺省汇总测量和分箱数。

- 单击该图标以打开**字段选择器**对话框并选择要添加到列表的字段。
- 在**聚集**列中，选择汇总测量。
- 对于连续字段，单击**分箱**列中的按钮以在**分箱**对话框中指定字段的定制分箱数。

空间时间预测

对于空间时间预测，在定义地图和数据源之后，向导中的其余步骤包括:

- 指定目标字段、时间字段和可选预测变量。
- 定义时间字段的时间间隔或循环周期。

根据需要，您可以:

- 选择不同输出选项。
- 定制模型构建参数。
- 定制聚集设置。
- 将预测值保存到当前会话中的数据集或保存为 IBM SPSS Statistics 格式的数据文件。

选择字段

可用字段列表包含来自所选数据源的字段。您可以通过从**数据源**列表选择数据源来控制所显示的字段列表。

目标 目标字段是必填字段。目标是要预测其值的字段。

- 目标字段必须是连续（刻度）的数字字段。
- 如果有两个数据源，那么目标是核心密度估算，并且“密度”显示为目标名称。您不能更改此选择。

预测变量

可以指定一个或多个预测变量字段。此设置是可选的。

时间字段

您必须选择一个或多个字段来表示时间段或选择**循环周期**。

- 如果有两个数据源，必须从两个数据源中选择时间字段。两个时间字段必须表示相同的时间间隔。
- 对于循环周期，您必须在向导的“时间间隔”面板中指定用于定义周期循环的字段。

时间间隔

此面板中的选项基于在选择字段步骤中所选的**时间字段**或**循环周期**。

时间字段

所选时间字段。如果您在选择字段步骤中选择了多个时间字段，那么这些字段将显示在此列表中。

时间间隔。从列表中选择相应时间间隔。根据时间间隔，您还可以指定其他设置，例如两次观察之间的间隔（增量）和开始值。此时间间隔用于选择的所有时间字段。

- 该过程假设所有个案（记录）都表示间隔相等的时间区间。

- 根据所选的时间间隔，该过程可以检测缺失的观察或者同一时间间隔中需要聚集在一起的多个观察。例如，如果时间间隔是天，日期 2014-10-27 后面是 2014-10-29，那么缺少 2014-10-28 的观察。如果时间间隔是月，那么同一个月份中的多个日期将聚集在一起。
- 对于某些时间间隔，条件设置可以定义间隔相等的正常时间区间中的断点。例如，如果时间间隔是天，但仅工作日有效，那么您可以指定一星期有五天，每星期从星期一开始。
- 如果所选时间字段不是日期格式或时间格式字段，那么时间间隔自动设置为**周期**，并且无法更改。

循环字段

如果您在选择字段步骤中选择**循环周期**，那么必须指定用于定义循环周期的字段。循环周期用于确定重复性循环变差，例如一年中的月份数或一星期中的天数。

- 您最多可以指定三个用于定义循环周期的字段。
- 第一个循环字段表示循环的最高级别。例如，如果有按年、季度和月份表示的循环变差，那么表示年的字段是第一个循环字段。
- 第一个和第二个循环字段的循环长度是后一个级别的周期。例如，如果循环字段是年、季度和月份，那么第一个循环长度是 4，第二个循环长度是 3。
- 第二个和第三个循环字段的开始值是其中每个循环字段的第一个值。
- 循环长度和开始值必须是正整数。

聚集

- 如果您在选择字段的步骤中选择任意**预测变量**，那么可以为预测变量选择聚集汇总方法。
- 如果定义的时间间隔中有多条记录，那么需要聚集。例如，如果时间间隔是月，那么同一个月份中的多个日期将聚集在一起。
- 您可以为连续和有序字段指定聚集汇总测量方法。名义字段基于最常见的值进行聚集。

连续 对于连续（刻度）字段，汇总测量可以是平均值、中间值或总和。

有序 对于有序字段，汇总测量可以是中间值、众数、最高值或最低值。

定制特定字段的设置

您可以覆盖特定预测变量的缺省聚集汇总测量。

- 单击该图标以打开**字段选择器**对话框并选择要添加到列表的字段。
- 在**聚集**列中，选择汇总测量。

输出

地图

目标值。

所选目标字段的值的地图。

相关性 相关性地图。

集群 突出显示彼此相似的位置集群的地图。

位置相似性阈值。

创建集群所需的相似性。该值必须是大于 0 且小于 1 的数字。

指定最大集群数。

要显示的最大集群数。

模型评估表

模型规范。
用于运行分析的规范摘要，包括目标、输入和位置字段。

时间信息摘要。
确定模型中使用的时间字段和时间间隔。

平均值结构效应检验。
输出包括模型和每个效应的检验统计值、自由度和显著性水平。

模型系数平均值结构。
输出包括每个模型项的系数值、标准误差、检验统计值、显著性水平和置信区间。

自回归系数。
输出包括每个延迟的系数值、标准误差、检验统计值、显著性水平和置信区间。

空间协方差检验。
对于基于方差图的参数模型，显示空间协方差结构的拟合优度测试结果。测试结果可确定是以参数方式对空间协方差结构建模，还是使用非参数方式模型。

参数空间协方差。
对于基于方差图的参数模型，显示参数空间协方差的参数估计。

模型选项

模型设置

自动包含截距
在模型中包含截距。

最大自回归
最大自回归。该值必须是介于 1 到 5 之间的整数。

空间协方差

指定空间协方差的估算方法。

参数化 估算方法是参数化方法。方法可以是**高斯**、**指数**或**幂指**。对于幂指，您可以指定**幂**值。

非参数化
估算方法是非参数化方法。

保存

将地图和上下文数据保存为地图规范

将地图规范保存为外部文件 (.mplan)。您可以将此地图规范文件上载到向导以供后续分析。您还可以将地图规范文件与 SPATIAL TEMPORAL PREDICTION 命令一起使用。

将任何地图和数据文件复制到规范
地图规范中所用的地图形状文件、外部数据文件和数据集中的数据保存在地图规范文件中。

评分 将目标字段的预测值、方差、置信度上限和下限保存在所选数据文件中。

- 您可以将预测值保存在当前会话中打开的数据集或保存为 IBM SPSS Statistics 格式的数据文件。
- 数据文件不能是模型中使用的数据源。
- 数据文件必须包含模型中使用的所有时间字段和预测变量。
- 时间值必须大于模型中使用的时间值。

高级

具有缺失值的最大个案数 (%)

具有缺失值的最大个案百分比。

显著性水平

显著性水平用于确定基于方差图的参数模型是否适当。该值必须大于 0 并小于 1。缺省值是 0.05。显著性水平用于空间协方差结构中的拟合优度测试。拟合优度统计量用于确定是使用参数化还是非参数化模型。

不确定性因子 (%)

不确定性因子是一个百分比值，表示未来预测的不确定性增长。随着进入将来的每个步骤，预测不确定性的上下限将按指定百分比增长。

完成

在“地理空间建模向导”的最后一步中，您可以运行模型，也可以将生成的命令语法粘贴到语法窗口。您可以修改和保存生成的语法以供后续使用。

声明

本信息是为在美国提供的产品和服务编写的。

IBM 可能在其他国家或地区不提供本文档中讨论的产品、服务或功能特性。有关您当前所在区域的产品和服务的信息，请向您当地的 IBM 代表咨询。任何对 IBM 产品、程序或服务的引用并非意在明示或暗示只能使用 IBM 的产品、程序或服务。只要不侵犯 IBM 的知识产权，任何同等功能的产品、程序或服务，都可以代替 IBM 产品、程序或服务。不过，用户应自行负责评估和验证任何非 IBM 产品、程序或服务的工作情况。

IBM 公司可能已拥有或正在申请与本文档内容有关的各项专利。您获得该文档并不意味着授予您任何这些专利许可。您可以将书面许可查询函件发送至:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

有关双字节 (DBCS) 信息的许可查询，请与您所在国家或地区的 IBM 知识产权部门联系，或用书面方式将查询寄往:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

本条款不适用英国或任何这样的条款与当地法律不一致的国家或地区: International Business Machines Corporation“按现状”提供本出版物，不附有任何种类的（无论是明示的还是暗含的）保证，包括但不限于暗含的有关非侵权、适销和适用于某种特定用途的保证。有些州/省不允许针对某些交易的明示或暗示免责条款，因此本声明可能不适用于您。

本信息可能含有技术误差或排版错误。此处的信息会定期进行更改；这些更改会体现在本出版物的新版本中。IBM 可以随时对本资料中描述的产品和/或程序进行改进和/或更改，而不另行通知。

本信息中对任何非 IBM Web 站点的引用都只是为了方便起见才提供的， 不以任何方式充当对那些 Web 站点的保证。那些 Web 站点中的资料不是 IBM 产品资料的一部分，使用那些 Web 站点带来的风险将由您自行承担。

IBM 可以按它认为适当的任何方式使用或分发您所提供的任何信息而无须对您承担任何责任。

本程序的被许可方如果要了解有关程序的信息以达到如下目的: （i）允许在独立创建的程序和其他程序（包括本程序）之间进行信息交换，以及（ii）允许对已经交换的信息进行相互使用，请与下列地址联系:

IBM Software Group
ATTN: Licensing

200 W. Madison St.
Chicago, IL; 60606
U.S.A.

此类信息的提供应遵照相关条款和条件，其中包括在某些情况下支付适当费用。

本资料中描述的许可程序及其所有可用的许可资料均由 IBM 依据 IBM 客户协议、IBM 国际软件许可协议或任何同等协议中的条款提供。

此处所含的性能数据均在受控环境下决定。因此，在其他操作环境中获得的结果可能差异较大。有些测量可能在开发级的系统中进行，不保证这些测量结果与常用系统上的测量结果相同。此外，有些测量结果可能通过推断来估计得出。实际结果可能有所差异。此文档的用户应针对其具体环境验证适用的数据。

非 IBM 产品的相关信息来自这些产品的供应商，及其发布的公告或其他公开来源。IBM 没有对这些产品进行测试，也无法确认其性能的精确性、兼容性或任何其他关于非 IBM 产品的声明。有关非 IBM 产品性能的问 题应当向这些产品的供应商提出。

有关 IBM 未来方向或意向的所有声明均可能未经通知即变更或撤销，并且仅代表目标和目的。

本信息包含日常业务运营中使用的数据和报告的示例。为了尽可能详尽地对其进行说明，示例中包含了人员的姓名、公司、品牌和产品的名称。所有这些名称均为虚构，与真实商业企业使用的名称和地址的任何雷同纯属巧合。

版权许可:

本信息包括源语言形式的样本应用程序，这些样本说明不同操作平台上的编程方法。如果是为按照在编写样本程序的操作平台上的应用程序编程接口 (API) 进行应用程序的开发、使用、经销或分发为目的，您可以任何形式对这些样本程序进行复制、修改、分发，而无须向 IBM 付费。这些示例并未在所有条件下作全面测试。因此，IBM 不能担保或暗示这些程序的可靠性、可维护性或功能。本样本程序仍然是“按现状”提供的，不附有任何种类的保证。对于因使用样本程序所引起的任何损害，IBM 概不负责。

凡这些实例程序的每份拷贝或其任何部分或任何衍生产品，都必须包括如下版权声明:

© （贵公司的名称）（年）。此部分代码是根据 IBM Corp. 公司的样本程序衍生出来的。

© Copyright IBM Corp. _（输入年份）_. All rights reserved.

商标

IBM、IBM 徽标和 ibm.com 是 International Business Machines Corp., 在全球许多管辖区域注册的商标或注册商标。其他产品和服务名称可能是 IBM 或其他公司的商标。当前的 IBM 商标列表，可从 Web 站点 www.ibm.com/legal/copytrade.shtml 上“版权和商标信息”部分获取。

Adobe、Adobe 徽标、PostScript 和 PostScript 徽标是 Adobe Systems Incorporated 在美国和/或其他国家或地区的注册商标或商标。

Intel、Intel 徽标、Intel Inside、Intel Inside 徽标、Intel Centrino、Intel Centrino 徽标、Celeron、Intel Xeon、Intel SpeedStep、Itanium 和 Pentium 是 Intel Corporation 或其子公司在美国和其他国家或地区的商标或注册商标。

Linux 是 Linus Torvalds 在美国、其他国家或地区或两者的注册商标。

Microsoft、Windows、Windows NT 和 Windows 徽标是 Microsoft Corporation 在美国、其他国家或地区或两者的商标。

UNIX 是 The Open Group 在美国和其他国家或地区的注册商标。

Java 和所有基于 Java 的商标和徽标是 Oracle 及/或其分支机构的商标和注册商标。

索引

[E]

[F]

[G]

[H]

[J]

[K]

[L]

[M]

[N]

[P]

[Q]

[R]

[Y]

[Z]

A

B

C

D

E

F

G

S

T

V

W

Y

Z

Printed in China