IBM SPSS Complex Samples 19

Note: Before using this information and the product it supports, read the general information under Notices第 259 页码.

前言

IBM® SPSS® Statistics 是一种用于分析数据的综合系统。复杂抽样 可选附加模块提供本手册中描述的其他分析方法。此 复杂抽样 附加模块必须与 SPSS Statistics Core 系统一起使用，并已完全集成到了该系统中。

关于 SPSS Inc.， IBM 下属公司

SPSS Inc. 是一家 IBM 下属公司，它也是全球领先的预测分析软件和解决方案提供商。该公司拥有全面的产品系列，涵盖数据收集、统计量、建模和部署，通过在业务流程中嵌入分析技术，收集人们的态度与看法，预测未来客户交互结果，然后针对这些深入见解采取相应行动。SPSS Inc. 解决方案着眼于整合分析技术、IT 基础设施和业务流程，以帮助达成整个企业内相互关联的业务目标。全球各地的众多企业、政府和学术机构客户依靠 SPSS Inc. 技术在吸引、留住和发展客户方面取得竞争优势，同时减少欺诈并缓解风险。SPSS Inc. 在 2009 年 10 月被 IBM 并购。有关更多信息，请访问 http://www.spss.com。

技术支持

我们提供有“技术支持”以维护客户。客户可就 SPSS Inc. 产品使用或某一受支持硬件环境的安装帮助寻求技术支持。要获得“技术支持”，请访问 SPSS Inc. 网站 http://support.spss.com，或通过网站 http://support.spss.com/default.asp?refpage=contactus.asp 找到当地办事处。在请求协助时，请准备好您和您组织的 ID 以及支持协议。

客户服务

如果对发货或帐户存在任何问题，请联系您当地的办事处，联系方式列在 Web 站点中，网址为 http://www.spss.com/worldwide。请先准备好您的序列号以供识别。

培训讲座

SPSS Inc. 提供公开的以及现场的培训讲座。所有讲座都是以实践小组为特色的。讲座将定期在各大城市开展。关于这些讲座的更多信息，请联系您本地的办事处，联系方式列在 Web 站点上，网址为 http://www.spss.com/worldwide。

附加出版物

SPSS Statistics：数据分析指南、SPSS Statistics：Statistical Procedures Companion 和 SPSS Statistics：Advanced Statistical Procedures Companion（由 Marija Norušis 编写，并已由 Prentice Hall 出版）作为建议的补充材料提供。这些出版物涵盖 SPSS Statistics Base 模块、Advanced Statistics 模块和 回归模块中的统计过程。无论您是刚开始从事数据分析工作，还是已准备好使用高级应用程序，这些书籍都将帮助您最有效地利用在 IBM® SPSS® Statistics 产品中找到的功能。有关其他信息，包括出版物的内容和示例章节，请参阅作者的网站： http://www.norusis.com

内容

部分 II: 示例

部分 I:
用户指南

章 1

复杂样本过程简介

在传统软件包中，一项对分析过程的固有假设是：数据文件中的观察数据代表从关注的群体选取的简单随机样本。这种假设对越来越多的公司不再适用，研究人员发现以更为结构化的方式获取样本既经济高效又很方便。

使用“复杂样本”选项，可以根据一项复杂设计选择样本，并将设计指定项融入数据分析中，从而确保结果是有效的。

复杂样本的属性

复杂样本在很多方面与简单随机样本不同。在简单随机样本中，各抽样单元是直接从整个总体中采用不放回方式以等概率 (WOR) 随机选择的。相比之下，给定的复杂样本具有以下部分或全部特征：

层次。分层抽样在总体的非重叠子组（即层次）中独立选择样本。例如，层次可以是社会经济组、工作类别、年龄组或种族组。通过分层，可以确保子组的样本大小足够大，提高整个估计值的精确度，并在不同层次使用不同抽样方法。

聚类。聚类抽样需要选择抽样单元组（即聚类）。例如，聚类可以是学校、医院或地理区域，抽样单元可以是学生、病人或市民。聚类在多阶段设计和区域（地理）样本中很常见。

多阶段。在多阶段抽样中，应基于聚类选择第一阶段样本。然后，通过从所选聚类抽取子样本创建第二阶段样本。如果第二阶段样本是基于子聚类的，则可以向样本添加第三阶段。例如，在调查的第一阶段，可以抽取城市样本。然后，从所选城市中，可以抽取家庭样本。最后，从所选家庭中，可以对个人进行民意调查。使用抽样和分析准备向导可以在一个设计中指定三个阶段。

非随机抽样。如果随机选择难以实现，则可以系统（以固定间隔）或顺序方式抽取单元。

不等选择概率。如果抽取的聚类包含的单元数不相等，可以使用与大小成正比 (PPS) 的概率进行抽样，以使聚类的选择概率与其所含单元的比例相等。PPS 抽样还可以使用更多一般加权设计来选择单元。

无限制抽样。无限制抽样以放回方式 (WR) 选择单元。因此，单个单元可能多次选入样本中。

抽样权重。抽样权重是在抽取复杂样本时自动计算的，与目标总体中每个抽样单元代表的“频率”十分一致。因此，根据样本的权重总和可以估计总体大小。复杂样本分析过程需要抽样权重以正确分析复杂样本。请注意：这些权重应该在“复杂样本”选项内使用，而不应通过“加权个案”过程用于其他分析过程，该过程将权重视为个案重复。

“复杂样本”过程的使用

“复杂样本”过程的使用取决于特定需要。用户主要类型为执行如下任务的人员：

- 根据复杂设计计划和执行调查，可能以后再分析样本。调查人员的主要工具是抽样向导。
- 根据复杂设计分析以前获得的样本数据文件。在使用“复杂样本”分析过程之前，可能需要使用分析准备向导。

无论是哪类用户，都需要向“复杂样本”过程提供设计信息。为便于重新使用，这一信息存储在**计划文件**中。

计划文件

规划文件包含复杂抽样规范。计划文件有两种类型：

抽样计划。抽样向导中给定的指定项定义用于抽取复杂样本的样本设计。抽样计划文件包含这些指定项。抽样计划文件还包含一个缺省分析计划，该计划使用适合指定样本设计的估计方法。

分析计划。此计划文件包含“复杂样本”分析过程正确计算复杂样本的方差估计所需的信息。该计划包括样本结构、每个阶段的估计方法和对所需变量（如样本权重）的引用。使用分析准备向导可以创建和编辑分析计划。

在计划文件中保存指定项有几个好处，包括：

- 调查人员可以指定多阶段抽样计划的第一阶段并立即抽取第一阶段单元，并为第二阶段收集抽样单元的信息，然后修改抽样计划以包括第二阶段。
- 不能访问抽样计划文件的分析人员可以指定一个分析计划，然后从每个“复杂样本”分析过程引用该计划。
- 大型公用样本的设计人员可以发布抽样计划文件，这样简化了分析人员指令，也使分析人员不再需要每人都指定自己的分析计划。

其他参考

有关抽样技术的更多信息，请参见以下内容：

Cochran, W. G. 1977. Sampling Techniques, 3rd ed. New York: John Wiley and Sons.

Kish, L. 1965. Survey Sampling. New York: John Wiley and Sons.

Kish, L. 1987. Statistical Design for Research. New York: John Wiley and Sons.

Murthy, M. N. 1967. Sampling Theory and Methods. Calcutta, India: Statistical Publishing Society.

Särndal, C., B. Swensson, 和 J. Wretman. 1992. Model Assisted Survey Sampling. New York: Springer-Verlag.

章
2

从复杂设计抽样

图片 2-1
抽样向导，“欢迎”步骤

该抽样向导将指导您完成创建、修改或执行抽样计划文件的步骤。在使用向导之前，应构思好定义明确的目标总体、抽样单元列表和适当的样本设计。

创建新样本计划

- 从菜单中选择：
  分析 > **复杂样本** > **选择样本...**
- 选择**设计样本**并选择一个计划文件名来保存样本计划。
- 单击**下一步**使向导继续。

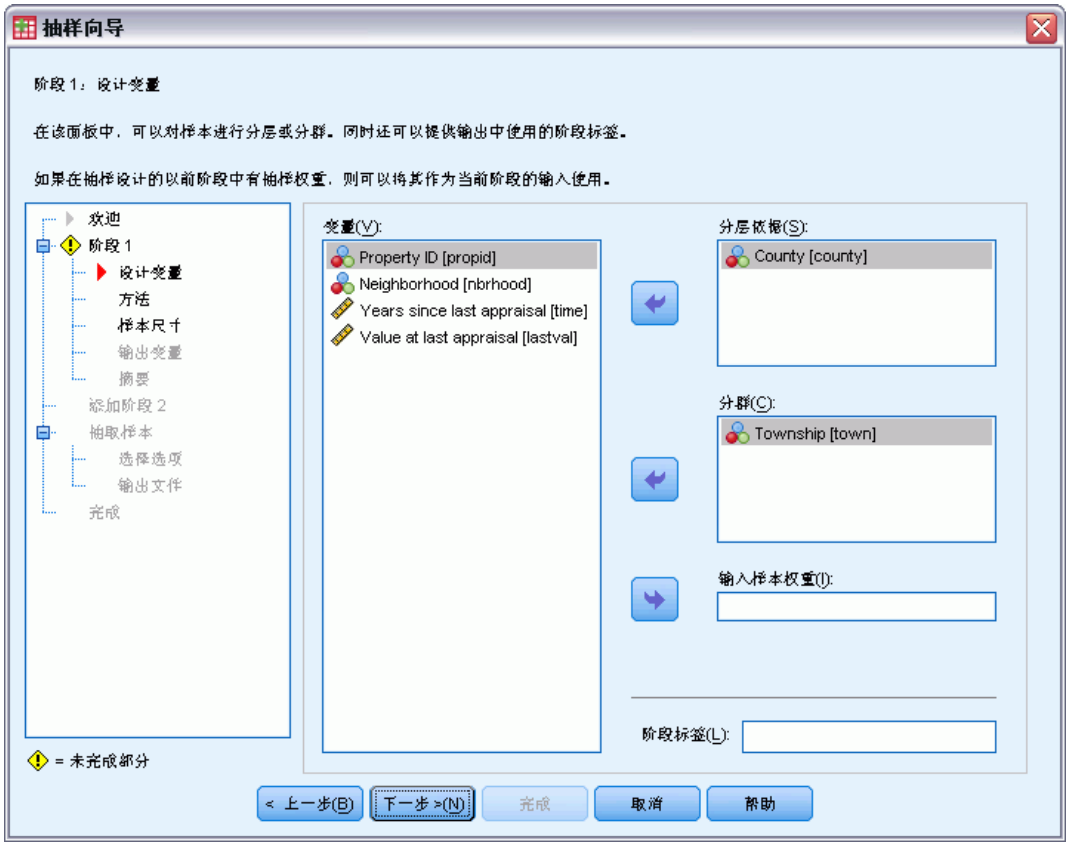
- 或者，在“设计变量”步骤中，可以定义层次、聚类和输入样本权重。定义这些内容之后，单击**下一步**。
- 或者，在“抽样方法”步骤中，可以选择一个方法用于选择样本。

  如果选择 PPS Brewer 或 PPS Murthy，可以单击**完成**抽取样本。否则，单击**下一步**，然后：
- 在“样本大小”步骤中，指定要抽样的单元数或单元比例。
- 现在，即可单击**完成**抽取样本。

  或者，可以进一步执行以下步骤：
  - 选择要保存的输出变量。
  - 向设计添加第二或第三阶段。
  - 设置各选择选项，包括抽取样本的阶段、随机数种子，以及是否将用户缺失值视为设计变量的有效值。
  - 选择输出数据的保存位置。
  - 将所选项粘贴为命令语法。

抽样向导：设计变量

图片 2-2
抽样向导，“设计变量”步骤



在这一步骤中，可以选择层次变量和聚类变量，可以定义输入样本权重。还可指定阶段的标签。

分层依据。分层变量的交叉分类定义了不同的子体，即层次。分别为各层获取了不同的样本。要提高估计值的精确度，层中单元的特征应尽量均一。

分群。分群变量定义观察单元组，即分群。如果从总体直接抽取观察单元很昂贵，或者不可能实现，就可以使用分群；可以从总体抽取分群，然后从所选分群抽取观察单元。但是，使用分群会在抽样单元之间引入相关性，导致精度下降。要使这种影响减到最小，分群中的单元的特征应尽量均一。必须至少定义一个分群变量才能计划多阶段设计。在使用多个不同抽样方法时，分群也是必不可少的。 有关详细信息，请参阅第 6 页码中的抽样向导：抽样方法.

输入样本权重。如果当前样本设计是更大样本设计的一部分，则可以从更大样本设计的以前阶段获得样本权重。在当前设计的第一阶段，可以指定一个包含这些权重的数值型变量。对于当前设计的后续阶段，样本权重将自动计算。

阶段标签。可为每个阶段指定一个可选的字符串标签。该标签用在输出中以帮助识别分阶段信息。

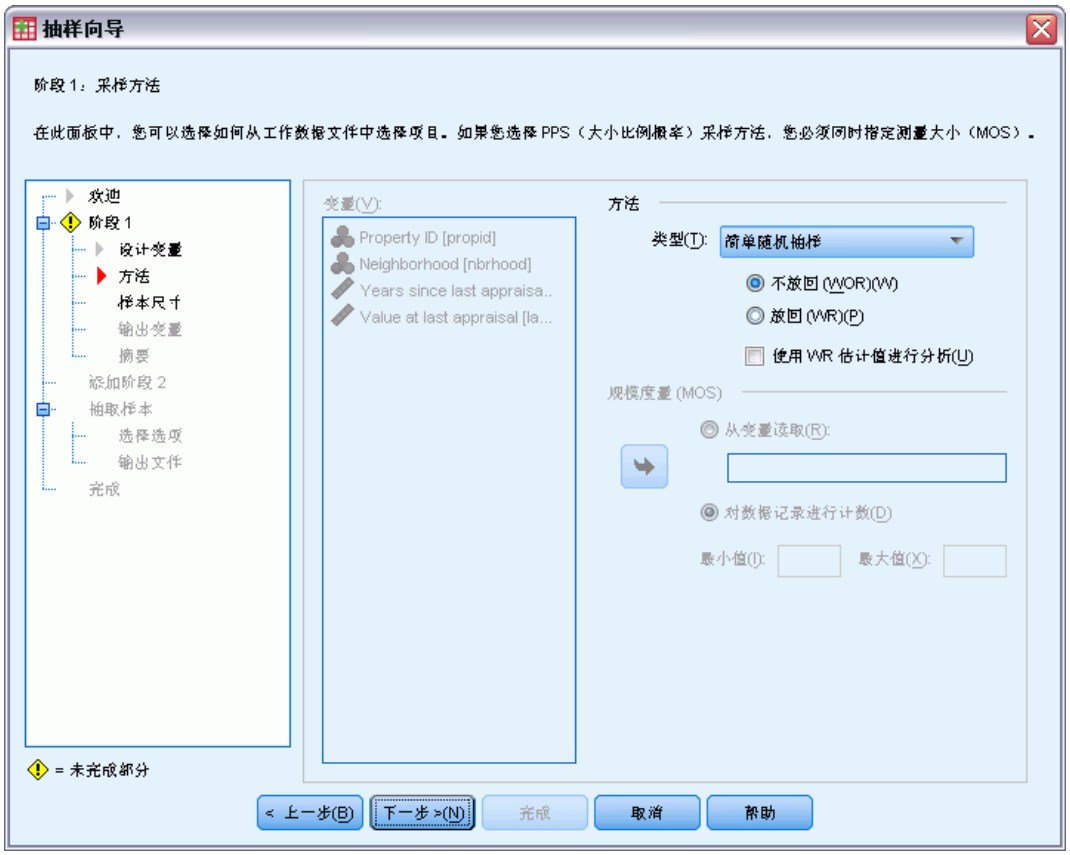
注意：源变量列表的内容在所有向导步骤中都相同。换言之，在某个特定步骤中从源列表移去的变量将在所有步骤中从该列表移去。返回源列表的变量在所有步骤中都会显示在列表中。

浏览抽样向导的树控件

在抽样向导的每个步骤中，左侧都是所有步骤的概要。单击概要中已启用步骤的名称可浏览该向导。只要前面的所有步骤有效—即前面每个步骤都具有要求的最小指定项，则步骤为启用状态。有关给定步骤无效原因的更多信息，请参见各步骤的“帮助”。

抽样向导：抽样方法

图片 2-3
抽样向导，“样本方法”步骤



在这一步骤中，可以指定从活动数据集中选择个案的方式。

方法。该组中的控件用于选择一种选择方法。某些抽样类型允许选择放回抽样 (WR) 或不放回抽样 (WOR)。有关更多信息，请参见类型描述。请注意，某些与大小成正比的概率 (PPS) 类型只在定义聚类之后才可用，所有 PPS 类型只在设计的第一阶段才可用。此外，WR 方法只在设计的最后阶段才可用。

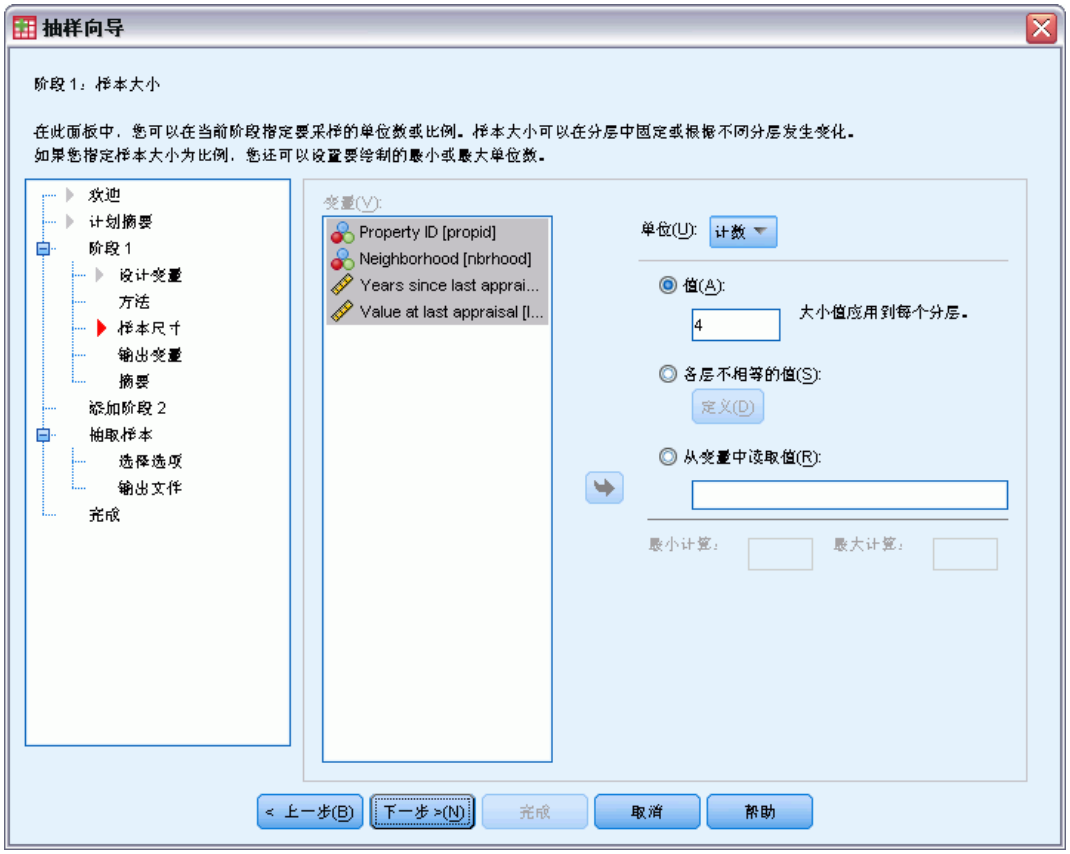
- **简单随机抽样。**以等概率选择单元。单元可以采用放回或不放回方式进行选择。
- **简单系统。**在整个抽样框架或层次（如果指定）中，采用不放回方式以固定间隔选择单元。在第一个区间内随机选择的单元即选作起始点。
- **简单顺序。**采用不放回方式以等概率顺序地选择单元。
- **PPS。**这是第一阶段方法，它以与大小成正比的概率随机选择单元。任何单元都可以采用放回方式选择；只有聚类可以采用不放回方式抽样。
- **PPS 系统。**这是第一阶段方法，它以与大小成正比的概率系统地选择单元。并且单元是以不放回方式选择的。
- **PPS 顺序。**这是第一阶段方法，它以与聚类大小成正比的概率采用不放回方式顺序选择单元。

- **PPS Brewer。**这是第一阶段方法，它以与聚类大小成正比的概率采用不放回方式从每个层次选择两个聚类。要使用此方法，必须指定聚类变量。
- **PPS Murthy。**这是第一阶段方法，它以与聚类大小成正比的概率采用不放回方式从每个层次选择两个聚类。要使用此方法，必须指定聚类变量。
- **PPS Sampford。**这是第一阶段方法，它以与聚类大小成正比的概率从每个层次采用不放回方式选择两个以上聚类。它是 Brewer 方法的扩展。要使用此方法，必须指定聚类变量。
- **在分析中使用 WR 估计。**缺省情况下，估计方法是在计划文件中指定的，与所选抽样方法一致。这样，即使抽样方法意味着 WOR 估计，也可以使用放回方式估计。此选项只在阶段 1 可用。

大小测量 (MOS)。如果选择 PPS 方法，则必须指定定义每个单元大小的规模度量。这些规模可以在一个变量中显式定义，也可以根据数据计算。或者，可以设置 MOS 的上限和下限，覆盖所有 MOS 变量中的值或根据数据计算的值。这些选项只在阶段 1 可用。

抽样向导：样本大小

图片 2-4
抽样向导，“样本大小”步骤



在这一步骤中，可以指定当前阶段中要抽样的单元数或单元比例。样本大小可以是固定的，也可以各层不同。为了指定样本大小，前面阶段中选择的聚类可用于定义层次。

单位。可以指定要抽样的单元的确切样本大小或比例。

- **值。**应用于所有层次的单个值。如果将**计数**选作单元度规，则应输入一个正整数。如果选择**比例**，则应输入一个非负值。除非是放回抽样，否则比例值也应不大于 1。
- **各层不相等的值。**允许您通过“定义不等大小”对话框逐层输入大小值。
- **从变量中读取值。**允许您选择包含层次大小值的数值变量。

如果选择**比例**，则可以设置抽样单元数的下限和上限。

定义不等大小

图片 2-5
“定义不等大小”对话框

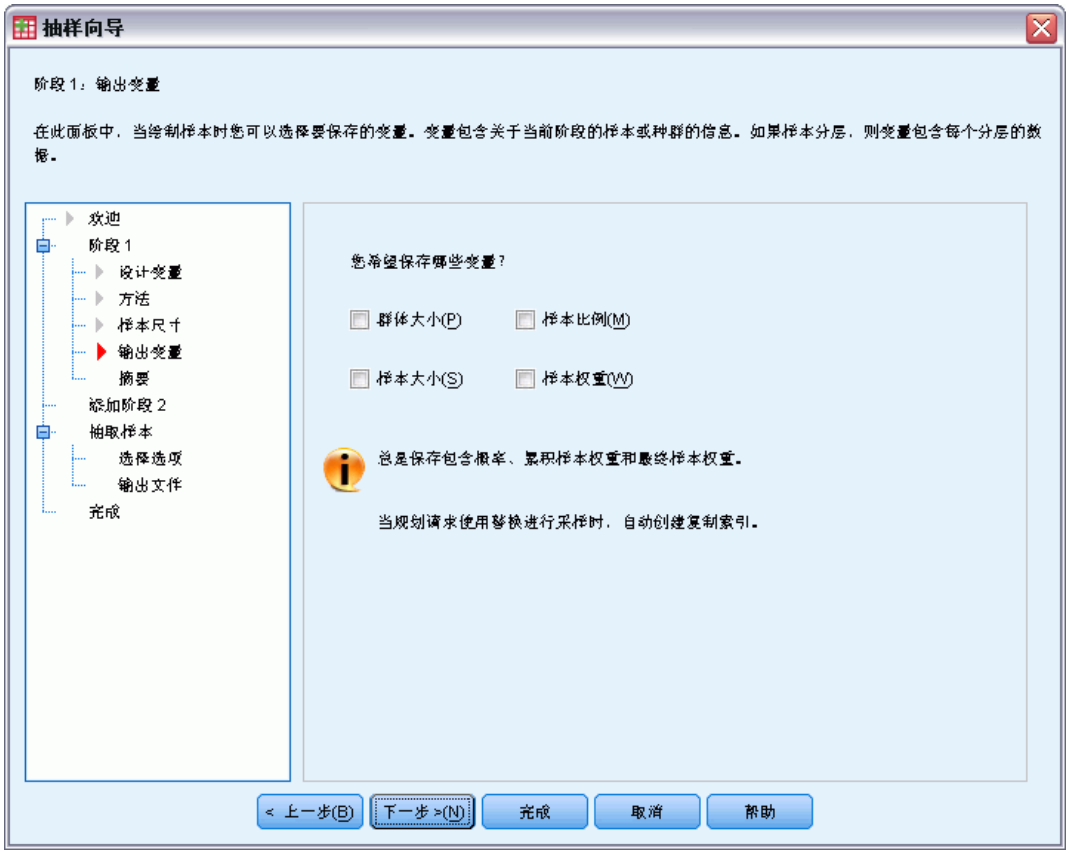
在“定义不等大小”对话框中，可以逐层输入大小值。

“指定大小”网格。该网格最多显示五个层次变量或聚类变量的交叉分类—即每行一个层次/聚类组合。符合的变量包括当前和以前阶段的所有分层变量，以及以前阶段的所有聚类变量。变量可在网格内重新排序，或者移到“排除”列表。在最右列中输入大小。单击**标签**或**值**，在网格单元格中分层变量和聚类变量的值标签和数据值的显示之间切换。包含未标注值的单元格始终显示值。单击**刷新层**，用网格中变量的标注数据值的每个组合重新填充网格。

排除。要指定层次/聚类组合子集的大小，请将一个或多个变量移到“排除”列表。这些变量不用于定义样本大小。

抽样向导：输出变量

图片 2-6
抽样向导，“输出变量”步骤



在这一步骤中，可以选择抽取样本时要保存的变量。

群体大小。给定阶段估计的总体单元数。保存变量的根名为 PopulationSize_。

样本比例。给定阶段的抽样率。保存变量的根名为 SamplingRate_。

样本大小。给定阶段抽取的单元数。保存变量的根名为 SampleSize_。

样本权重。包含概率的逆。保存变量的根名为 SampleWeight_。

某些分阶段变量是自动生成的。其中包括：

包含概率。给定阶段抽取的单元比例。保存变量的根名为 InclusionProbability_。

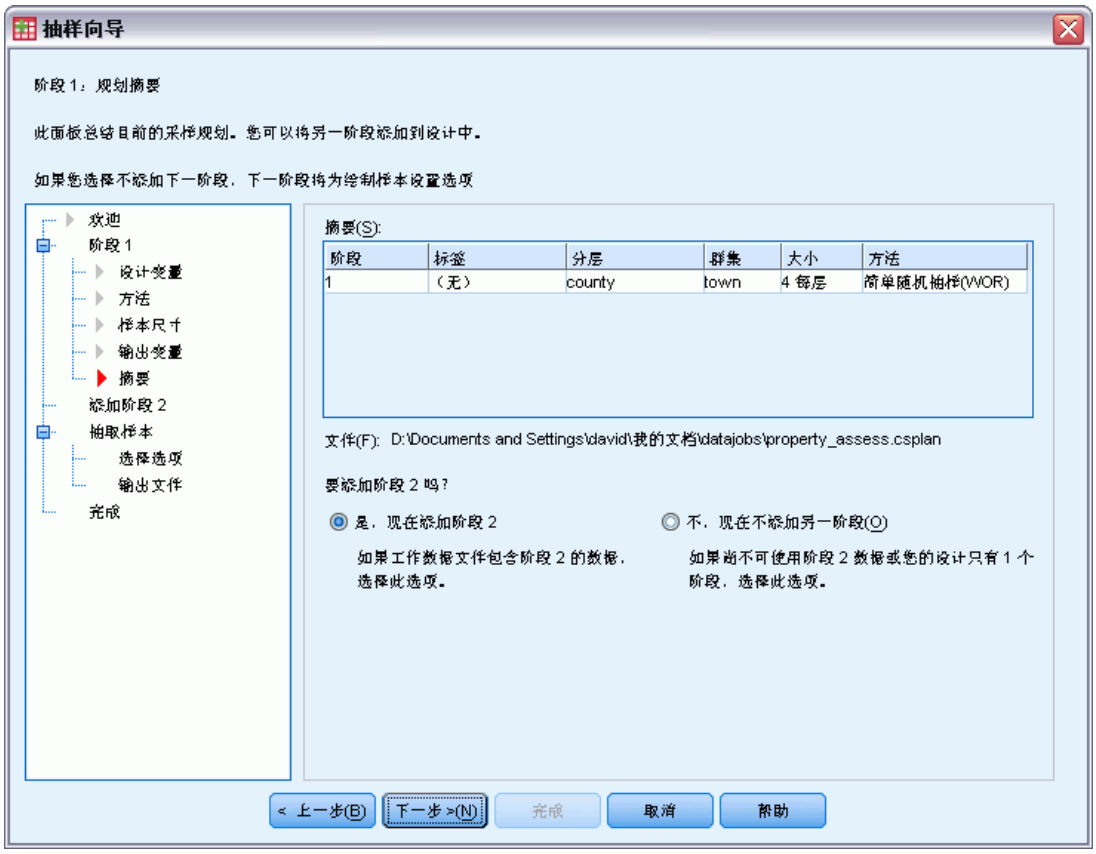
累积权重。当前阶段之前（包括当前阶段）的累积样本权重。保存变量的根名为 SampleWeightCumulative_。

指标。标识给定阶段中多次选择的单元。保存变量的根名为 Index_。

注意：保存变量根名称包含一个整数后缀，它反映阶段号—例如，PopulationSize_1_ 用于阶段 1 保存的总体大小。

抽样向导：计划摘要

图片 2-7
抽样向导，“计划摘要”步骤



这是每个阶段的最后一步，提供整个当前阶段的样本设计指定项的摘要。在此，可以继续下一阶段（必要时创建阶段），也可以设置抽取样本的选项。

抽样向导：抽取样本：选择选项

图片 2-8
抽样向导，“抽取样本选择选项”步骤

在这一步骤中，可以选择是否抽取样本。还可以控制其他抽样选项，如随机种子和缺失值处理。

抽取样本。除了选择是否抽取样本之外，还可以选择执行部分抽样设计。必须按顺序抽取阶段—即抽取阶段 1 之后才能抽取阶段 2。编辑或执行计划时，不能重新抽取锁定的阶段。

种子。它可用来选择生成随机数的种子值。

包括用户缺失值。可确定用户缺失值是否有效。如果有效，则将用户缺失值视为单独的类别。

数据已排序。如果样本框架按照分层变量值预先排序，使用此选项可以加快选择过程。

抽样向导：抽取样本：输出文件

图片 2-9
抽样向导，抽取样本，“输出文件”步骤

在这一步骤中，可以选择抽样个案、权重变量、联合概率和个案选择规则的定向位置。

样本数据。使用这些选项可以确定样本输出的写入位置。样本输出可以添加到活动数据集，也可以写入新数据集，还可以保存到外部 IBM® SPSS® Statistics 数据文件中。数据集在当前会话期间可用，但在后续会话期间不可用，除非显式将其保存为数据文件。数据集名称必须符合变量命名规则。 如果指定外部文件或新数据集，则写入所选个案的抽样输出变量和活动数据集中的变量。

联合概率。使用这些选项可以确定联合概率的写入位置。联合概率保存到外部 SPSS Statistics 数据文件。如果选择 PPS WOR、PPS Brewer、PPS Sampford 或 PPS Murthy 方法，并且未指定 WR 估计，则会生成联合概率。

个案选择规则。如果要一次在一个阶段构造样本，可能希望将个案选择规则保存到文本文件中。个案选择规则对于构造后续阶段的子框架非常有用。

抽样向导：完成

图片 2-10
抽样向导，“完成”步骤

这是最后一步。现在即可保存计划文件并抽取样本，或者将选择内容粘贴到语法窗口中。

在对现有计划文件中的阶段进行更改时，可将已编辑的计划另存为新文件或覆盖现有文件。如果添加阶段而不更改现有阶段，则向导将自动覆盖现有计划文件。如果要将计划保存为新文件，请选择**将向导生成的语法粘贴到语法窗口**，并在语法命令中更改文件名。

修改现有样本计划

- 从菜单中选择：
  分析 > **复杂样本** > **选择样本**...
- 选择**编辑样本设计**，并选择要编辑的计划文件。
- 单击**下一步**使向导继续。
- 在“计划摘要”步骤中复查抽样计划，然后单击**下一步**。

  后续步骤与新设计大体相同。有关更多信息，请参见各步骤的“帮助”。
- 浏览到“完成”步骤，为编辑过的计划文件指定新名称，或选择覆盖现有计划文件。

根据需要，您可以：

- 指定已进行抽样的阶段。
- 从计划中移去阶段。

抽样向导：计划摘要

图片 2-11
抽样向导，“计划摘要”步骤

在这一步骤中，可以复查抽样计划，确定已进行了抽样的阶段。如果编辑计划，还可以从计划中移去阶段。

以前抽样的阶段。如果扩展抽样框架不可用，则必须一次在一个阶段执行多阶段抽样设计。从下拉列表中选择哪些阶段已进行了抽样。所有执行过的阶段都是锁定的；它们在“抽取样本选择选项”步骤中不可用，并且不能在编辑计划时更改。

移去阶段。可从多阶段设计中移去阶段 2 和 3。

运行现有样本计划

▶ 从菜单中选择：
分析 > 复杂样本 > 选择样本...

- 选择**抽取样本**，并选择要运行的计划文件。
- 单击**下一步**使向导继续。
- 在“计划摘要”步骤中复查抽样计划，然后单击**下一步**。
- 执行样本计划时，会跳过包含阶段信息的各步骤。现在，随时可以执行“完成”步骤。根据需要，您可以指定已进行抽样的阶段。

CSPLAN 和 CSSELECT 命令附加功能

使用命令语法语言还可以：

- 为输出变量指定定制名称。
- 在浏览器中控制输出。例如，如果设计或修改了样本，则可以取消显示所显示的分阶段计划摘要，如果执行了样本设计，则可以取消显示所显示的按层抽样个案的分布摘要，并请求显示个案处理摘要。
- 在活动数据集中选择一个变量子集，以写入外部样本文件或其他数据集。

请参见命令语法参考以获取完整的语法信息。

章
3

准备复杂样本以进行分析

图片 3-1
分析准备向导，“欢迎”步骤

分析准备向导将引导您完成创建或修改分析计划的各个步骤，以用于各种“复杂样本”分析过程。使用该向导之前，应先根据一项复杂设计完成样本抽取。

如果不能访问用于抽取样本的抽样计划文件（该抽样计划包含一个缺省分析计划），则创建一个新的计划非常有用。如果确实可以访问用于抽取样本的抽样计划文件，则可以使用抽样计划文件包含的缺省分析计划，也可以覆盖缺省分析指定项并将更改保存到新文件中。

创建新的分析计划

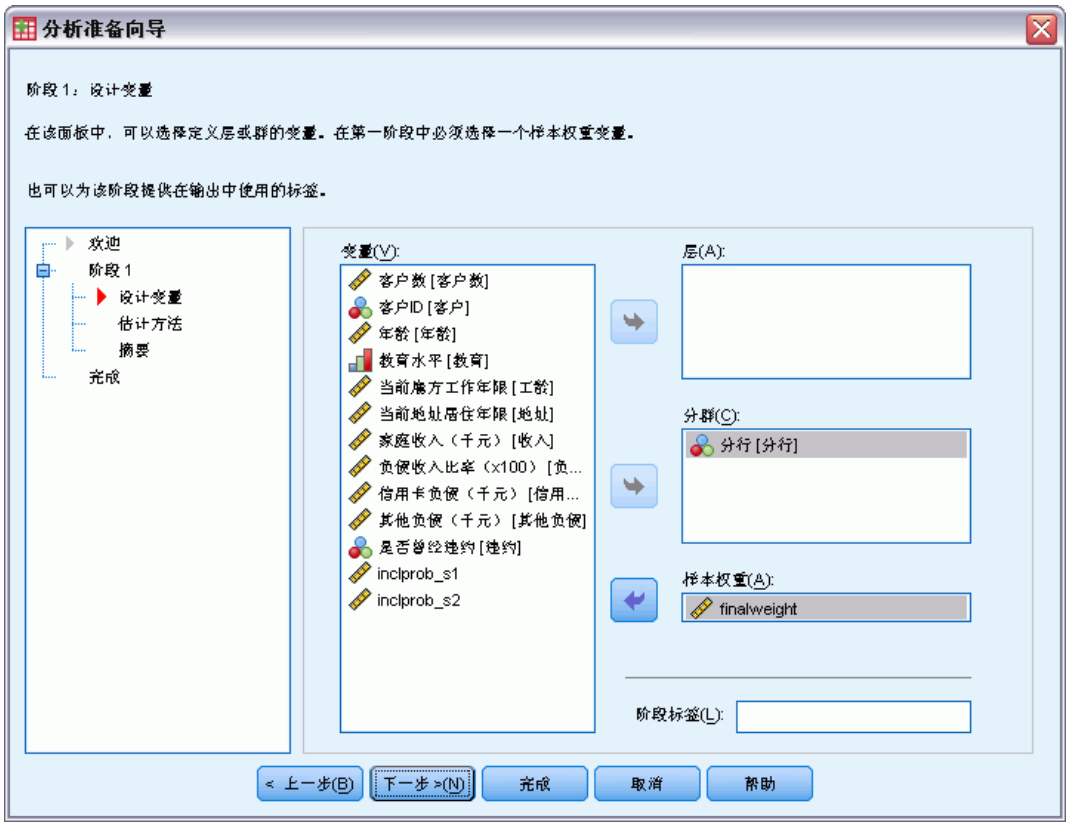
▶ 从菜单中选择：
分析 > 复杂样本 > 准备分析...

- 选择**创建计划文件**，然后选择一个计划文件名，用于保存分析计划。
- 单击**下一步**使向导继续。
- 在“设计变量”步骤中，指定包含样本权重的变量，定义层次和聚类（可选）。
- 现在，即可单击**完成**保存计划。

  或者，可以进一步执行以下步骤：
  - 在“估计方法”步骤中，选择用于估计标准误的方法。
  - 在“大小”步骤中，指定抽取的单元数或每个单元的包含概率。
  - 向设计添加第二或第三阶段。
  - 将所选项粘贴为命令语法。

分析准备向导：设计变量

图片 3-2
分析准备向导，“设计变量”步骤



在这一步骤中，可以确定层变量和分群变量并定义样本权重。还可指定阶段的标签。

层。分层变量的交叉分类定义了不同的子体，即层次。总样本代表每层的独立样本的组合。

分群。分群变量定义观察单元组，即分群。多阶段抽取的样本首先在较早阶段中选择分群，然后从所选分群中抽取子样本单元。在分析通过放回方式分群抽样获得的数据文件时，应将重复指数包括为分群变量。

样本权重。必须在第一阶段提供样本权重。对于当前设计的后续阶段，样本权重将自动计算。

阶段标签。可为每个阶段指定一个可选的字符串标签。该标签用在输出中以帮助识别分阶段信息。

注意：源变量列表的内容在该向导所有步骤中都相同。换言之，在某个特定步骤中从源列表移去的变量将在所有步骤中从该列表移去。会在所有步骤中显示返回源列表的变量。

浏览分析向导的树控件

在每个分析向导步骤的左侧都是所有步骤的概要。单击概要中已启用步骤的名称可浏览该向导。只要前面所有步骤都有效—即只要该步骤前面每个步骤都具有要求的最小指定项，则步骤为启用状态。有关给定步骤无效原因的更多信息，请参见各步骤的“帮助”。

分析准备向导：估计方法

图片 3-3
分析准备向导，“估计方法”步骤

在这一步骤中，可以指定阶段的估计方法。

WR（放回式抽样）。在复杂抽样设计下估计方差时，WR 估计不包括对有限总体抽样的修正 (FPC)。在简单随机抽样 (SRS) 下估计方差时，可以选择包括或排除 FPC。

如果分析权重已进行标度，建议选择不包括用于 SRS 方差估计的 FPC，以免分析权重增加总体大小。SRS 方差估计用于计算类似于设计效果的统计量。只能在设计的最后阶段指定 WR 估计；如果选择 WR 估计，向导将不允许添加其他阶段。

等概率 WOR（等概率不放回式抽样）。等概率 WOR 估计包括有限总体修正，并假设单元是等概率抽取的。等概率 WOR 可在设计的任何阶段指定。

不等概率 WOR（不等概率不放回式抽样）。除了使用有限总体修正之外，不等概率 WOR 还考虑以不等概率选择的抽样单元（通常为聚类）。此估计方法仅在第一阶段可用。

分析准备向导：字体大小

图片 3-4
分析准备向导，“大小”步骤

这一步骤用于指定当前阶段的包含概率或总体大小。大小可以是固定的，也可以各层不同。为了指定大小，前面阶段中指定的聚类可用于定义层次。请注意：仅当选择等概率 WOR 作为估计方法时，这一步骤才是必需的。

单位。可以指定精确的总体大小或单元抽样概率。

- **值。**应用于所有层次的单个值。如果将**总体大小**选作单元度规，则应输入一个非负整数。如果选择**包含概率**，则应输入一个 0 到 1 之间（包括 0 和 1）的值。

- **各层不相等的值。**允许您通过“定义不等大小”对话框逐层输入大小值。
- **从变量中读取值。**允许您选择包含层次大小值的数值变量。

定义不等大小

图片 3-5
“定义不等大小”对话框

在“定义不等大小”对话框中，可以逐层输入大小值。

“指定大小”网格。该网格最多显示五个层次变量或聚类变量的交叉分类—即每行一个层次/聚类组合。符合的变量包括当前和以前阶段的所有分层变量，以及以前阶段的所有聚类变量。变量可在网格内重新排序，或者移到“排除”列表。在最右列中输入大小。单击**标签**或**值**，在网格单元格中分层变量和聚类变量的值标签和数据值的显示之间切换。包含未标注值的单元格始终显示值。单击**刷新层**，用网格中变量的标注数据值的每个组合重新填充网格。

排除。要指定层次/聚类组合子集的大小，请将一个或多个变量移到“排除”列表。这些变量不用于定义样本大小。

分析准备向导：计划摘要

图片 3-6
分析准备向导，“计划摘要”步骤

这是每个阶段的最后一个步骤，提供整个当前阶段的分析设计指定项的摘要。在此，可以继续下一阶段（必要时创建下一阶段），也可以保存分析指定项。

如果不能添加其他阶段，可能的原因如下：

- 在“设计变量”步骤中未指定任何聚类变量。
- 在“估计方法”步骤中选择了 WR 估计。
- 这是分析的第三阶段，而该向导最多支持三个阶段。

分析准备向导：完成

图片 3-7
分析准备向导，“完成”步骤

这是最后一步。现在，可以保存计划文件，或将选择内容粘贴到语法窗口。

在对现有计划文件中的阶段进行更改时，可将已编辑的计划另存为新文件或覆盖现有文件。如果添加阶段而不更改现有阶段，则向导将自动覆盖现有计划文件。如果要将计划保存为新文件，请选择**将本向导生成的语法粘贴到语法窗口**，并在语法命令中更改文件名。

修改现有分析计划

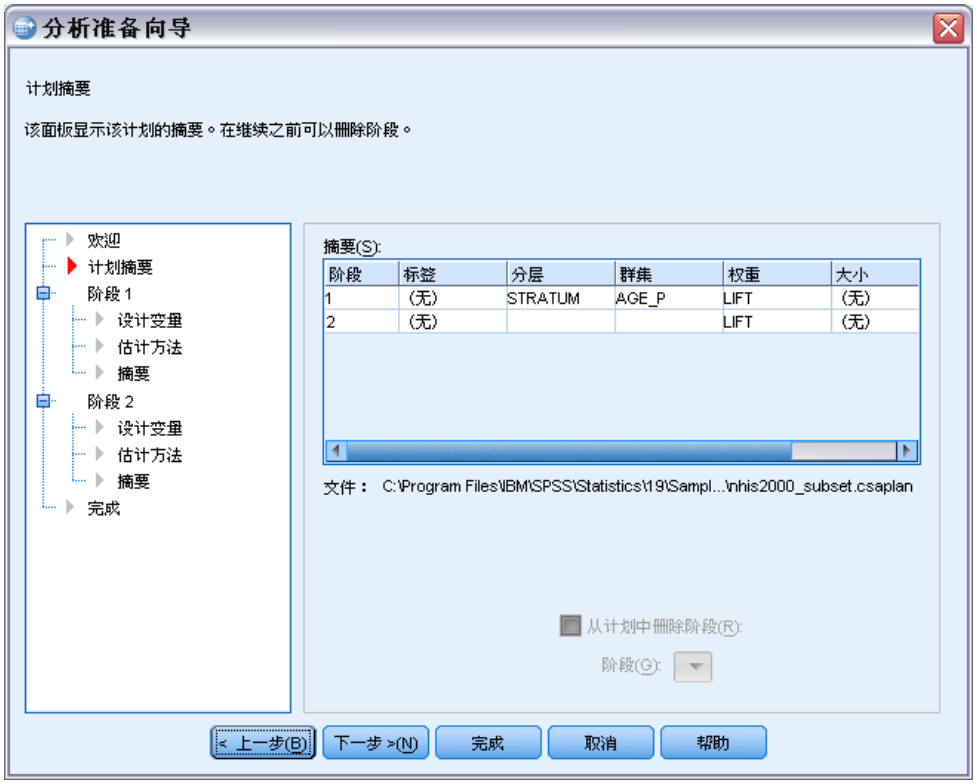
- 从菜单中选择：
  分析 > **复杂样本** > **准备分析...**
- 选择**编辑计划文件**，然后选择一个计划文件名，用于保存分析计划。
- 单击**下一步**使向导继续。
- 在“计划摘要”步骤中复查分析计划，然后单击**下一步**。

  后续步骤与新设计大体相同。有关更多信息，请参见各步骤的“帮助”。
- 浏览到“完成”步骤，为编辑过的计划文件指定新名称，或选择覆盖现有计划文件。

  根据需要，您可以从计划中删除阶段。

分析准备向导：计划摘要

图片 3-8
分析准备向导，“计划摘要”步骤



在这一步骤中，可以复查分析计划和从计划中移去阶段。

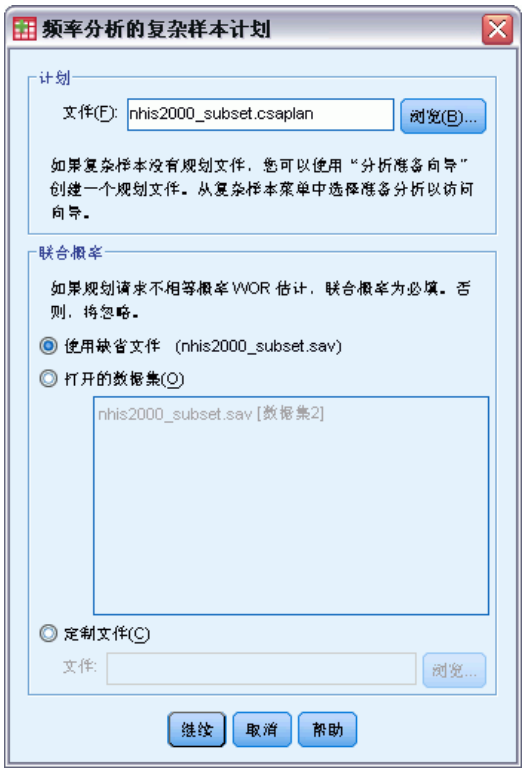
移去阶段。可从多阶段设计中移去阶段 2 和 3。计划必须至少有一个阶段，因此，可以编辑阶段 1，但不能将其从设计中移去。

章
4

复杂样本计划

“复杂样本”分析过程从分析或样本计划文件中获得分析指定项，以便提供有效的结果。

图片 4-1
“复杂样本计划”对话框



计划。指定分析或样本计划文件的路径。

联合概率。要对用 PPS WOR 方法抽取的聚类使用不等概率 WOR 估计，需要指定包含联合概率的单独文件或打开的数据集。此文件或数据集由抽样向导在抽样过程中创建。

章
5

复杂样本频率

“复杂样本频率”过程可以为所选变量生成频率表并显示单变量统计。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

示例。使用“复杂样本频率”过程，基于全美国健康访问调查 (NHIS) 的结果和这一公用数据的适当分析计划，可以获得美国公民维生素使用情况的单变量制表统计量。

统计量。该过程生成单元总体大小和表百分比的估计值，以及每个估计值的标准误、置信区间、变异系数、设计效果、设计效果平方根、累计值和未加权的计数。此外，还计算等单元比例检验的卡方和似然比统计量。

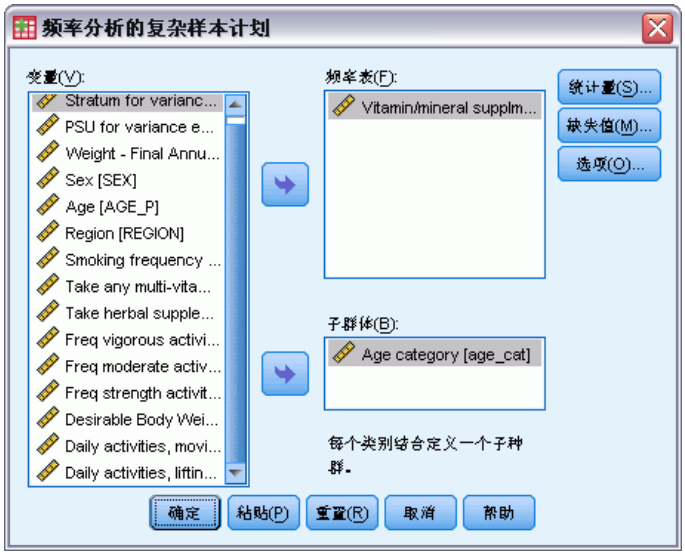
数据。要为其生成频率表的变量应为分类变量。子体变量可以是字符串或数值，但应该是分类变量。

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

获取复杂样本频率

- 从菜单中选择：
  分析 > **复杂样本** > **频率...**
- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

图片 5-1
“频率”对话框



- 选择至少一个频率变量。

根据需要，您可以指定变量来定义子体。统计量是针对每个子体分别计算的。

复杂样本频率：统计量

图片 5-2
“频率：统计量”对话框

单元格。在这一组中，可以请求单元格总体大小和表百分比的估计值。

统计量。这一组生成与总体大小或表百分比关联的统计量。

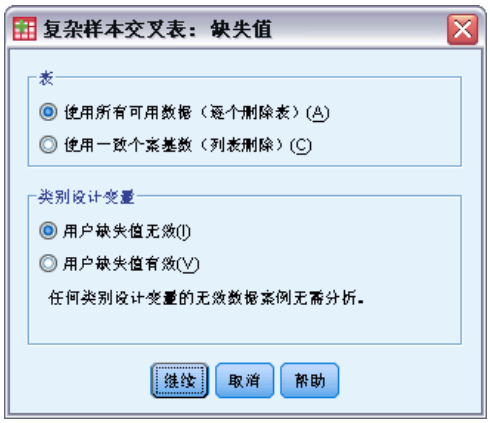
- **标准误。**估计值的标准误。
- **置信区间。**估计值的置信区间，使用指定水平。

- **变异系数。**估计值的标准误对估计值的比率。
- **去权重计数。**用于计算估计值的单元数。
- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**是指定复杂设计的效果的测量值，值与 1 相差越大表示效果越好。
- **累加值。**通过变量的每个值获得的累计估计值。

等单元格比例检验。对某个类别的变量频率相等的假设生成卡方和似然比检验。对每个变量进行独立检验。

复杂样本：缺失值

图片 5-3
“缺失值”对话框



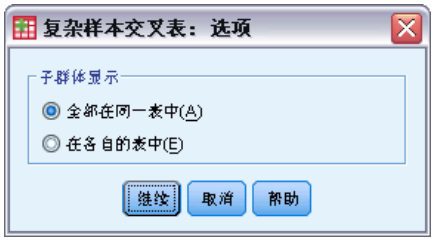
表。此组确定在分析中使用的个案。

- **使用所有可用数据。**缺失值是按照逐个表的方式确定的。因此，在频率表或交叉制表之间，用于计算统计量的个案可能不同。
- **确保一致个案基数。**缺失值是跨所有变量确定的。因此，用于计算统计量的个案在所有表中都一致。

类别设计变量。此组确定用户缺失值是否有效。

复杂样本：选项

图片 5-4
“选项”对话框



子群体显示。可以选择将子群体显示在同一个表或不同的表中。

章
6

复杂样本描述

“复杂样本描述”过程为多个变量显示单变量摘要统计量。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

示例。 使用“复杂样本描述”过程，基于全美国健康访问调查 (NHIS) 的结果和这一公用数据的适当分析计划，可以获得美国公民活动水平的单变量描述统计量。

统计量。该过程生成均值和总和，以及每个估计值的 t 检验、标准误、置信区间、变异系数、未加权的计数、总体大小、设计效果和设计效果平方根。

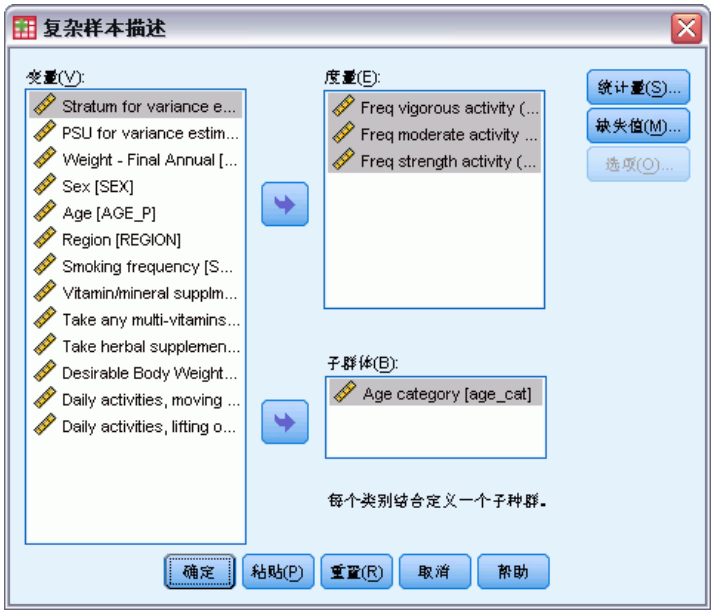
数据。测量应为尺度变量。子体变量可以是字符串或数值，但应该是分类变量。

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

获取复杂样本描述

- 从菜单中选择：
  分析 > 复杂样本 > 描述...
- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

图片 6-1
“描述”对话框



- 选择至少一个测量变量。

根据需要，您可以指定变量来定义子体。统计量是针对每个子体分别计算的。

复杂样本描述：统计量

图片 6-2
“描述：统计量”对话框

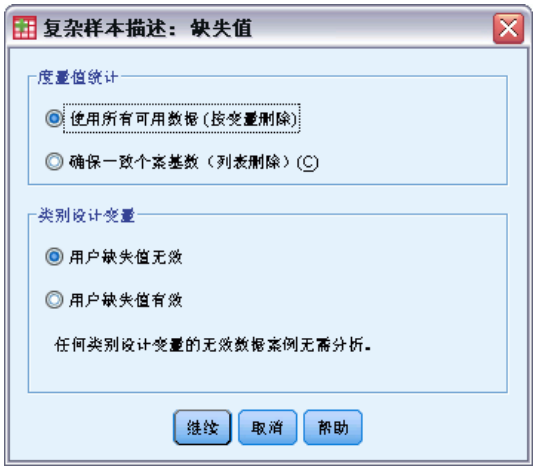
摘要。在这一组中，可以请求测量变量的均值和总和的估计值。此外，还可以请求针对指定值进行估计值的 t 检验。

统计量。这一组生成与均值或总和关联的统计量。

- **标准误。**估计值的标准误。
- **置信区间。**估计值的置信区间，使用指定水平。
- **变异系数。**估计值的标准误对估计值的比率。
- **去权重计数。**用于计算估计值的单元数。
- **群体大小。**总体中估计的单元数。
- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**是指定复杂设计的效果的测量值，值与 1 相差越大表示效果越好。

复杂样本描述：缺失值

图片 6-3
“描述：缺失值”对话框



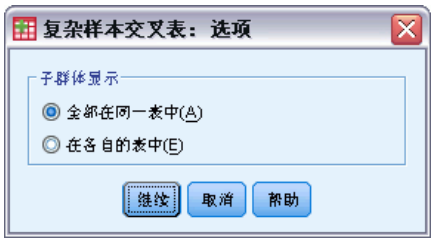
度量值统计。此组确定在分析中使用的个案。

- **使用所有可用数据。**缺失值是逐个变量确定的，因此，各测量变量用于计算统计量的个案可能不同。
- **确保一致个案基数。**缺失值是通过所有变量确定的，因此，用于计算统计量的个案是一致的。

类别设计变量。此组确定用户缺失值是否有效。

复杂样本：选项

图片 6-4
“选项”对话框



子群体显示。可以选择将子群体显示在同一个表或不同的表中。

章
7

复杂样本交叉表

复杂样本交叉表过程可以为所选变量对生成交叉表并显示二阶统计量。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

示例。使用“复杂样本交叉表”过程，基于全美国健康访问调查 (NHIS) 的结果和这一公用数据的适当分析计划，可以获得美国公民维生素使用量和抽烟频率的交叉分类统计量。

统计量。该过程生成单元格总体大小、行百分比、列百分比和表百分比的估计值，以及每个估计值的标准误、置信区间、变异系数、期望值、设计效果、设计效果平方根、残差、调整的残差和未加权的计数。计算几率比、相对风险和危险度差值以在 2x2 表中使用。此外，还计算 Pearson 和似然比统计量用于行变量和列变量的独立性检验。

数据。行变量和列变量应是分类变量。子体变量可以是字符串或数值，但应该是分类变量。

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

获取复杂样本交叉表

- 从菜单中选择：
  分析 > 复杂样本 > 交叉表...
- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

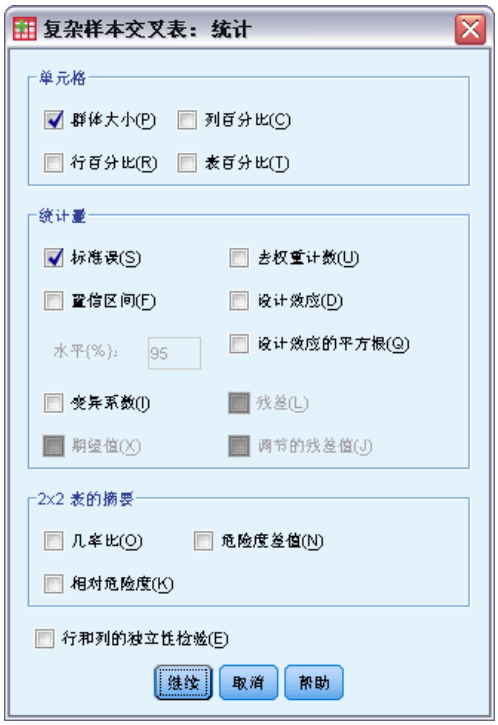
图片 7-1
“交叉表”对话框

- 选择至少一个行变量和一个列变量。

  根据需要，您可以指定变量来定义子体。统计量是针对每个子体分别计算的。

复杂样本交叉表：统计量

图片 7-2
“交叉表：统计量”对话框



单元格。在这一组中，可以请求单元格总体大小、行百分比、列百分比和表百分比的估计值。

统计量。这一组生成与总体大小、行百分比、列百分比和表百分比关联的统计量。

- **标准误。**估计值的标准误。
- **置信区间。**估计值的置信区间，使用指定水平。
- **变异系数。**估计值的标准误对估计值的比率。
- **期望值。**在假设行变量和列变量独立的条件下，估计值的期望值。
- **去权重计数。**用于计算估计值的单元数。
- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**是指定复杂设计的效果的测量值，值与 1 相差越大表示效果越好。
- **残差。**如果两个变量之间没有关系，则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，则正的残差表示单元中的实际个案数多于期望的个案数。
- **调整的残差。**单元格的残差（观察值减去期望值）除以其标准误的估计值。生成的标准化残差表示为均值上下的标准差单位。

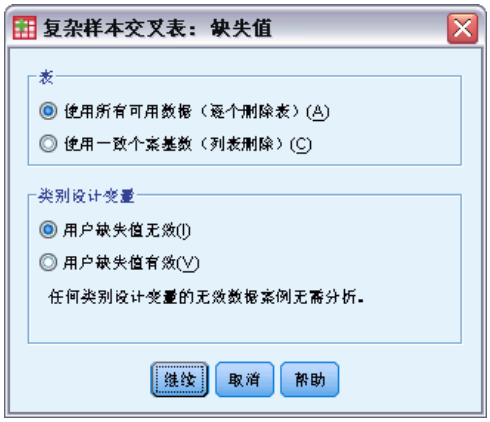
2x2 表的摘要。这一组为每个行变量和列变量都具有两个类别的表生成统计量。每一个都是某因子的存在与某事件的发生之间相关性大小的测量。

- **几率比。**当因子很少出现时，几率比可用作相对风险的估计值。
- **相对危险度。**存在因子出现事件的风险与不存在因子出现事件的风险的比率。
- **危险度差值。**存在因子出现事件的风险与不存在因子出现事件的风险之差。

行和列的独立性检验。生成行变量和列变量独立的假设的卡方检验和似然比检验。对每对变量进行单独检验。

复杂样本：缺失值

图片 7-3
“缺失值”对话框



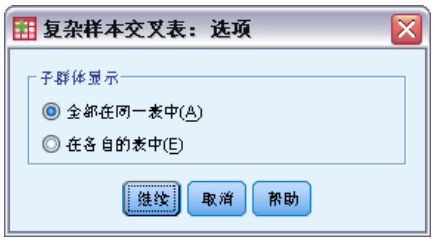
表。此组确定在分析中使用的个案。

- **使用所有可用数据。**缺失值是按照逐个表的方式确定的。因此，在频率表或交叉制表之间，用于计算统计量的个案可能不同。
- **确保一致个案基数。**缺失值是跨所有变量确定的。因此，用于计算统计量的个案在所有表中都一致。

类别设计变量。此组确定用户缺失值是否有效。

复杂样本：选项

图片 7-4
“选项”对话框



子群体显示。可以选择将子群体显示在同一个表或不同的表中。

章
8

复杂样本比率

“复杂样本比率”过程显示变量的比率的单变量摘要统计。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

示例。使用“复杂样本比率”过程，基于全国范围调查（根据一项复杂设计并采用适合数据的分析计划进行）的结果，可以获取当前财产价值与上次评估价值的比率的描述统计量。

统计量。该过程生成比率估计值、t 检验、标准误、置信区间、变异系数、未加权的计数、总体大小、设计效果和设计效果平方根。

数据。分子和分母应为正值刻度变量。子体变量可以是字符串或数值，但应该是分类变量。

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

获取复杂样本比率

- 从菜单中选择：
  分析 > 复杂样本 > 比率...
- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

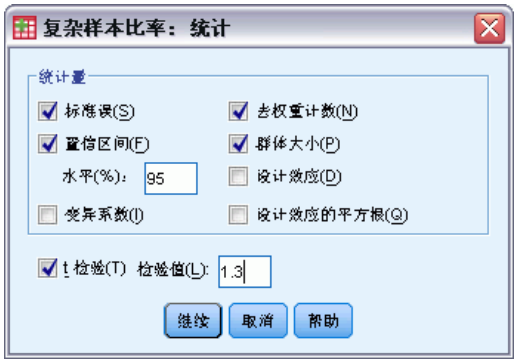
图片 8-1
“比率”对话框

- 选择至少一个分子变量和一个分母变量。

根据需要，您可以指定变量来定义要为其生成统计量的子组。

复杂样本比率：统计量

图片 8-2
“比率：统计量”对话框



统计量。此组生成与比率估计值关联的统计量。

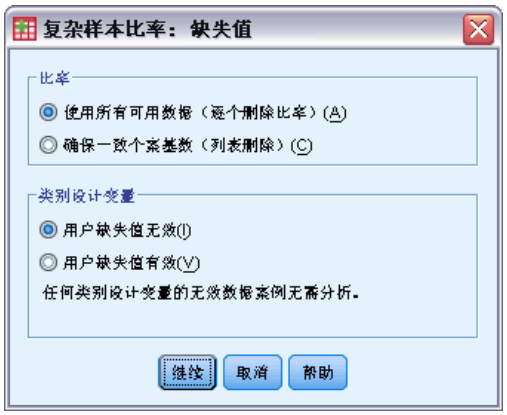
- **标准误。**估计值的标准误。
- **置信区间。**估计值的置信区间，使用指定水平。
- **变异系数。**估计值的标准误对估计值的比率。
- **去权重计数。**用于计算估计值的单元数。
- **群体大小。**总体中估计的单元数。

- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**是指定复杂设计的效果的测量值，值与 1 相差越大表示效果越好。

T 检验。可以请求针对指定值进行估计值的 t 检验。

复杂样本比率：缺失值

图片 8-3
“比率：缺失值”对话框



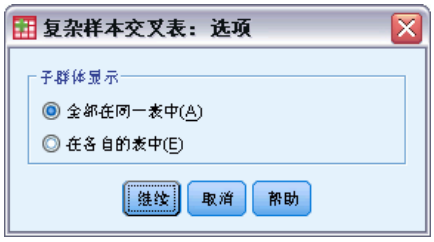
比率。此组确定在分析中使用的个案。

- **使用所有可用数据。**缺失值是按照逐个比率的方式确定的。因此，用于计算统计值的个案在各个分子分母对间可能不同。
- **确保一致个案基数。**缺失值是跨所有变量确定的。因此，用于计算统计量的个案是一致的。

类别设计变量。此组确定用户缺失值是否有效。

复杂样本：选项

图片 8-4
“选项”对话框



子群体显示。可以选择将子群体显示在同一个表或不同的表中。

章
9

复杂样本一般线性模型

“复杂样本一般线性模型”(CSGLM) 过程对通过复杂抽样方法抽取的样本执行线性回归分析以及方差和协方差分析。您还可以请求对子体进行分析。

示例。根据一项复杂设计，杂货连锁店对一组顾客的物习惯进行调查。在获得了调查结果以及每个顾客在上个月的消费金额之后，商店希望了解顾客购物的频率是否与他们在一个月中的消费金额有关，从而针对顾客性别进行控制并采用抽样设计。

统计量。该过程生成模型参数的估计值、标准误、置信区间、t 检验、设计效果和设计效果平方根，以及参数估计值之间的相关系数和协方差；还可以生成模型拟合的测量和自变量、因变量的描述统计量。此外，您还可以请求模型因子和因子交互的水平的估计边际均值。

数据。因变量是定量变量。因子是分类变量。协变量是与因变量相关的定量变量。子体变量可以是字符串或数值，但应该是分类变量。

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

获取复杂样本一般线性模型

从菜单中选择：

分析 > 复杂抽样 > 一般线性模型...

- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

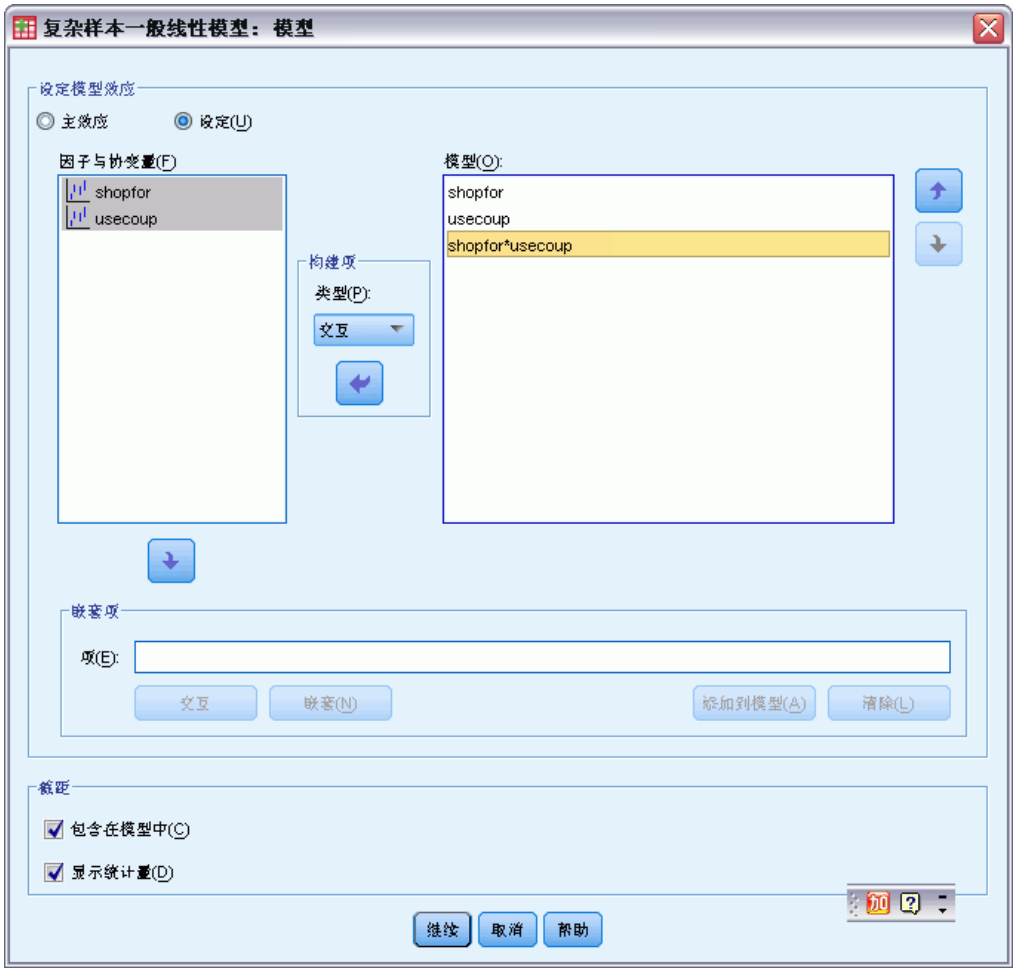
图片 9-1
“一般线性模型”对话框

▶ 选择一个因变量。

根据需要，您可以：

- 为因子和协变量选择适合您数据的变量。
- 指定用于定义子体的变量。仅对子体变量的所选类别执行该分析。

图片 9-2
“模型”对话框



指定模型效应。缺省情况下，该过程使用主对话框中指定的因子和协变量构建主效应模型。此外，还可以构建包含交互效应和嵌套项的定制模型。

非嵌套项

对于选定因子和协变量：

交互。为所有选定变量创建最高级交互项。

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二阶交互。

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。

嵌套项

在此过程中，可为您的模型建立嵌套项。嵌套项有助于对其值不与另一个因子的水平交互作用的因子或协变量的效应进行建模。例如，杂货连锁店可能在不同商店位置迎合顾客的不同消费习惯。由于每位顾客只频繁光顾某一位置的商店，因此 Customer 效应可以说是**嵌套在** Store location 效应中。

此外，还可以包含交互效应，例如包含相同协变量的多项式项，或将多层嵌套添加到嵌套项。

限制。嵌套项有以下限制：

- 一次交互内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A*A 是无效的。
- 嵌套效应内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A(A) 是无效的。
- 效应不可嵌套在协变量中。因此，如果 A 是因子且 X 是协变量，则指定 A(X) 是无效的。

截距。模型中通常包含截距。如果您可以假设数据穿过原点，则可以排除截距。即使在模型中包含截距，也可以选择取消显示与之相关的统计量。

复杂样本一般线性模型：统计量

图片 9-3
“一般线性模型：统计量”对话框

模型参数。使用此组可以控制与模型参数有关的统计量的显示。

- **估算。**显示系数的估计值。
- **标准误。**显示每个系数估计值的标准误。
- **置信区间。**显示每个系数估计值的置信区间。在“选项”对话框中设置该区间的置信度。
- **T 检验。**显示每个系数估计值的 t 检验。每个检验的原假设是该系数的值为 0。
- **参数估值协方差。**显示模型系数的协方差矩阵的估计值。
- **参数估值的相关性。**显示模型系数的相关性矩阵的估计值。

- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**是指定复杂设计的效果的测量值，值与 1 相差越大表示效果越好。

模型拟合。显示 R^2 和根均方误差统计量。

因变量和协变量的群体均值。显示有关因变量、协变量和因子的摘要信息。

样本设计信息。显示有关样本的摘要信息，包括未加权的计数和总体大小。

复杂样本假设检验

图片 9-4
“假设检验”对话框

检验统计。在这一组中，可以选择用于检验假设的统计类型。可以在 F、调整的 F、卡方和调整的卡方之间选择。

样本自由度。在这一组中，可以控制用于计算所有检验统计量的 p 值的抽样设计自由度。如果基于抽样设计，该值为抽样第一阶段的主抽样单元数和层数之差。或者，也可以通过指定一个正整数设置定制自由度。

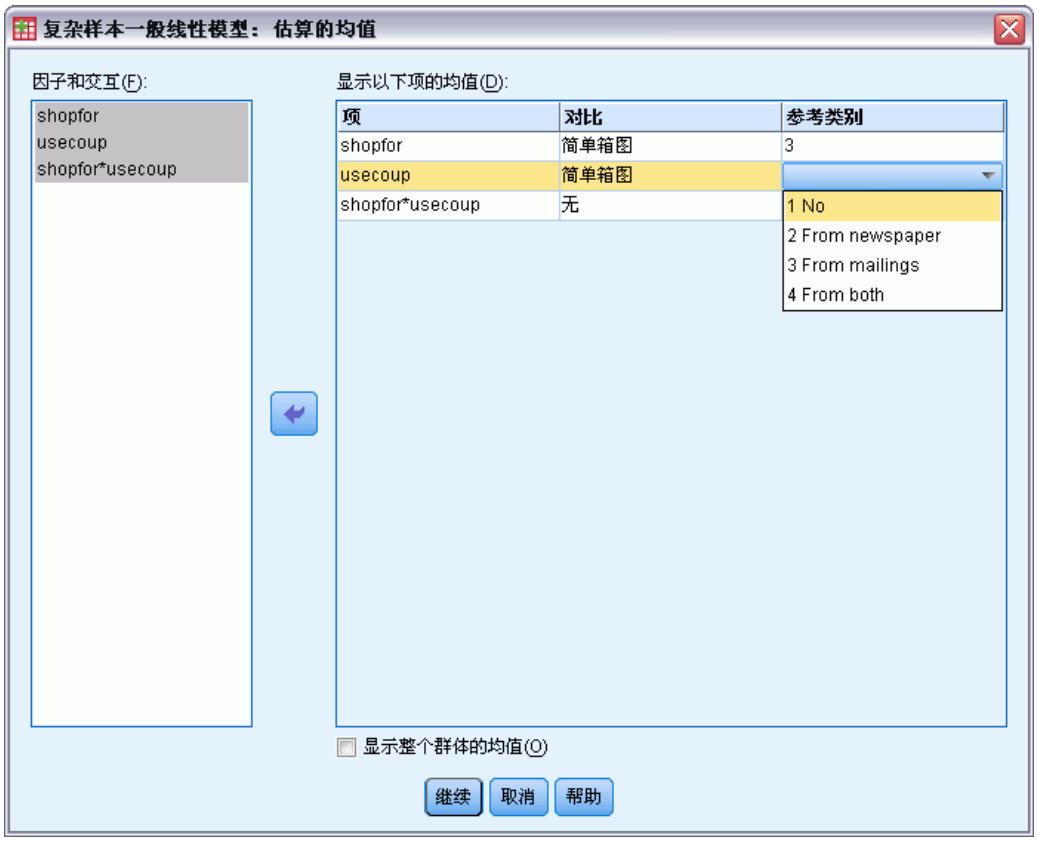
调整的多重比较。在执行包含多重比较的假设检验时，总体显著性水平可从所包含的比较的显著性水平进行调节。使用此组可以选择调节方法。

- **显著性最低的差异。**此方法并不控制拒绝某些线性对比不同于原假设值这一假设的总体概率。
- **连续 Sidak.** 这是按顺序逐步降低的拒绝 Sidak 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。
- **连续 Bonferroni.** 这是按顺序逐步降低的拒绝 Bonferroni 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。

- **Sidak**. 此方法提供比 Bonferroni 方法更严密的界限。
- **Bonferroni**. 此方法针对检验多个对比这一事实调整观测的显著性水平。

复杂样本一般线性模型：估算的均值

图片 9-5
“一般线性模型：估算的均值”对话框



在“估算的均值”对话框中，可以控制在“模型”子对话框中指定的因子和因子交互水平的模型估计边际均值的显示。还可以请求显示整个总体的均值。

项。计算所选因子和因子交互的估计均值。

对比。对比确定设置假设检验以比较估计均值的方式。

- **简单箱图**. 将每个水平的均值与指定水平的均值进行比较。当存在控制组时，此类对比很有用。
- **偏差**. 将每个水平（参考类别除外）的均值与所有水平的均值（总均值）进行比较。因子的水平可以为任何顺序。
- **差值**. 将每个水平（第一个除外）的均值与先前水平的均值进行比较。有时候将其称为逆 Helmert 对比。
- **Helmert**. 将因子的每个水平的均值（最后一个水平除外）与后续水平的均值进行比较。

- **重复**. 将每个水平的均值（最后一个水平除外）与后续水平的均值进行比较。
- **多项式**. 比较线性作用、二次作用、三次作用等等。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。

参考类别。简单对比和偏移对比需要参考类别或与其他因子水平进行比较的因子水平。

复杂样本一般线性模型：保存

图片 9-6
“一般线性模型：保存”对话框

保存变量。在这一组中，可以将模型预测值和残差保存为工作文件中的新变量。

将模型导出为 SPSS Statistics 数据。写入一个 IBM® SPSS® Statistics 格式的数据集，包含具有参数估计值、标准误、显著性值和自由度的参数相关性或协方差矩阵。矩阵文件中变量顺序如下。

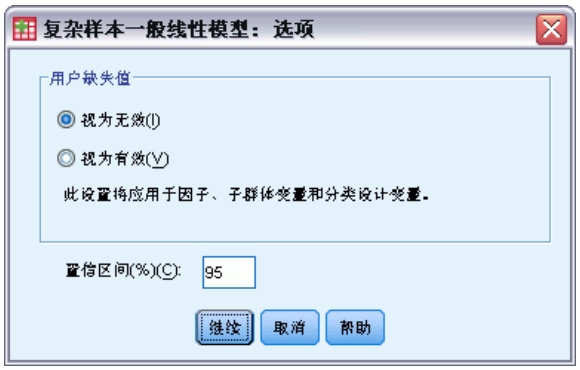
- **RowType_。**取值（或值标签）为 COV（协方差）、CORR（相关性）、EST（参数估计）、SE（标准误）、SIG（显著性水平）和 DF（抽样设计自由度）。存在每个模型参数的 COV（或 CORR）行类型的单独个案，以及每个其他行类型的个案。
- **VarName_。**对于行类型 COV 或 CORR，取值为 P1、P2、...，对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串。对于其他行类型，单元格为空。
- **P1、P2、...**这些变量对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串，这些变量根据行类型取值。对于冗余参数，所有协方差设为零；相关性设为系统缺失值；所有参数估计值设为零；并且所有标准误、显著性水平和残差自由度设为系统缺失值。

注意：该文件不能立即用于在其他读取矩阵文件的过程中执行进一步分析，除非这些过程接受在此导出的所有行类型。

将模型导出为 XML。将参数估计值和参数协方差矩阵（如果选择）以 XML (PMML) 格式保存。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。

复杂样本一般线性模型：选项

图片 9-7
“一般线性模型：选项”对话框



用户缺失值。所有设计变量以及因变量和任何协变量都必须包含有效数据。从分析中删除含有任何这些变量的无效数据的个案。使用这些控制可以确定是否在层次、聚类、子体和因子变量中将用户缺失值视为有效。

置信区间。这是系数估值和估计边际均值的置信区间水平。指定大于等于 50 且小于 100 的值。

CSGLM 命令附加功能

使用命令语法语言还可以：

- 指定作用对比作用的线性组合或一个值的自定义检验（使用 CUSTOM 子命令）。
- 在计算估计边际均值时修正值不同于其均值的协变量（使用 EMMEANS 子命令）。
- 为多项式对比指定度规（使用 EMMEANS 子命令）。
- 指定用于检查奇异性的容差值（使用 CRITERIA 子命令）。
- 为保存的变量创建用户指定名称（使用 SAVE 子命令）。
- 生成常规可估计函数表（使用 PRINT 子命令）。

请参阅命令语法参考以获取完整的语法信息。

章
10

复杂样本 Logistic 回归

“复杂样本 Logistic 回归”过程对通过复杂抽样方法抽取的样本的二元或多项因变量执行 logistic 回归分析。您还可以请求对子体进行分析。

示例。根据一项复杂设计，信贷员收集了在几个不同分支机构贷款的客户的过去记录。融入样本设计时，信贷员希望了解客户拖欠的概率是否与年龄、工作经历和信用负债量有关。

统计量。该过程生成模型参数的估计值、取幂估值、标准误、置信区间、t 检验、设计效果和设计效果平方根，以及参数估计值之间的相关系数和协方差。还可以生成因变量和自变量的伪 R^2 统计量、分类表和描述统计量。

数据。因变量是分类变量。因子是分类变量。协变量是与因变量相关的定量变量。子体变量可以是字符串或数值，但应该是分类变量。

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

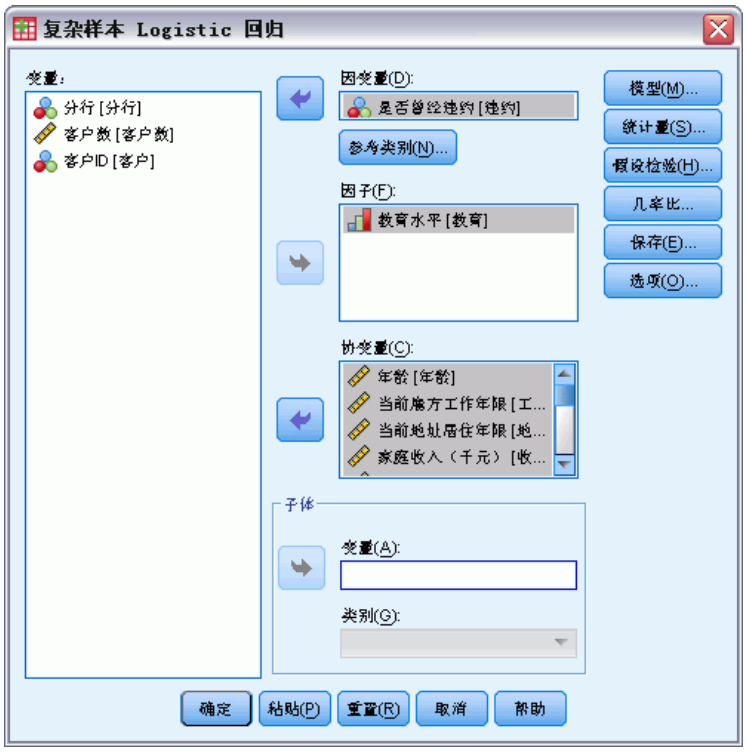
获取复杂样本 Logistic 回归

从菜单中选择：

分析 > **复杂抽样** > **Logistic 回归...**

- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

图片 10-1
“Logistic 回归”对话框



- 选择一个因变量。

根据需要，您可以：

- 为因子和协变量选择适合您数据的变量。
- 指定用于定义子体的变量。仅对子体变量的所选类别执行该分析。

复杂样本 Logistic 回归：参考类别

图片 10-2
“Logistic 回归：参考类别”对话框

默认情况下，“复杂样本 Logistic 回归”过程将最高值类别作为参考类别。在此对话框中，可以将最高值、最低值或自定义类别指定为参考类别。

复杂样本 Logistic 回归：模型

图片 10-3
“Logistic 回归：模型”对话框

指定模型效应。缺省情况下，该过程使用主对话框中指定的因子和协变量构建主效应模型。此外，还可以构建包含交互效应和嵌套项的定制模型。

非嵌套项

对于选定因子和协变量：

交互。为所有选定变量创建最高级交互项。

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二阶交互。

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。

嵌套项

在此过程中，可为您的模型建立嵌套项。嵌套项有助于对其值不与另一个因子的水平交互作用的因子或协变量的效应进行建模。例如，杂货连锁店可能在不同商店位置迎合顾客的不同消费习惯。由于每位顾客只频繁光顾某一位置的商店，因此 Customer 效应可以说是**嵌套在** Store location 效应中。

此外，还可以包含交互效应，例如包含相同协变量的多项式项，或将多层嵌套添加到嵌套项。

限制。嵌套项有以下限制：

- 一次交互内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A*A 是无效的。
- 嵌套效应内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A(A) 是无效的。
- 效应不可嵌套在协变量中。因此，如果 A 是因子且 X 是协变量，则指定 A(X) 是无效的。

截距。模型中通常包含截距。如果您可以假设数据穿过原点，则可以排除截距。即使在模型中包含截距，也可以选择取消显示与之相关的统计量。

复杂样本 Logistic 回归：统计量

图片 10-4
“Logistic 回归：统计量”对话框

模型拟合度。控制度量总体模型性能的统计量的显示。

- **伪 R 方。**在 Logistic 回归模型中，没有与线性回归的 R^2 完全对应的统计量。相反，却有多种试图模拟 R^2 统计量的属性的测量。
- **分类表。**按因变量的模型预测类别显示观察类别的制表交叉分类。

参数。使用此组可以控制与模型参数有关的统计量的显示。

- **估算。**显示系数的估计值。
- **取幂估值。**显示以系数估值为幂的自然对数的底数。当该估值对于统计检验有良好的属性时，取幂估值（即 exp(B)）更易于解释。
- **标准误。**显示每个系数估计值的标准误。
- **置信区间。**显示每个系数估计值的置信区间。在“选项”对话框中设置该区间的置信度。
- **T 检验。**显示每个系数估计值的 t 检验。每个检验的原假设是该系数的值为 0。
- **参数估值协方差。**显示模型系数的协方差矩阵的估计值。
- **参数估值的相关性。**显示模型系数的相关性矩阵的估计值。
- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**是指定复杂设计的效果的测量值，值与 1 相差越大表示效果越好。

模型变量的摘要统计。显示有关因变量、协变量和因子的摘要信息。

样本设计信息。显示有关样本的摘要信息，包括未加权的计数和总体大小。

复杂样本假设检验

图片 10-5
“假设检验”对话框

检验统计。在这一组中，可以选择用于检验假设的统计类型。可以在 F、调整的 F、卡方和调整的卡方之间选择。

样本自由度。在这一组中，可以控制用于计算所有检验统计量的 p 值的抽样设计自由度。如果基于抽样设计，该值为抽样第一阶段的主抽样单元数和层数之差。或者，也可以通过指定一个正整数设置定制自由度。

调整的多重比较。在执行包含多重比较的假设检验时，总体显著性水平可从所包含的比较的显著性水平进行调节。使用此组可以选择调节方法。

- **显著性最低的差异。**此方法并不控制拒绝某些线性对比不同于原假设值这一假设的总体概率。
- **连续 Sidak.** 这是按顺序逐步降低的拒绝 Sidak 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。
- **连续 Bonferroni.** 这是按顺序逐步降低的拒绝 Bonferroni 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。
- **Sidak.** 此方法提供比 Bonferroni 方法更严密的界限。
- **Bonferroni.** 此方法针对检验多个对比这一事实调整观测的显著性水平。

复杂样本 Logistic 回归：几率比

图片 10-6
“Logistic 回归：几率比”对话框

使用“几率比”对话框，可以控制指定的因子和协变量的模型估计几率比的显示。计算因变量的每个类别（参考类别除外）的一组单独几率比。

因子。对于所选的每个因子，都显示该因子处于每个类别的几率与处于指定参考类别的几率之比。

协变量。对于所选的每个协变量，都显示协变量均值加上指定变化单位的几率与处于均值的几率之比。

在计算因子或协变量的几率比时，该过程将修正所有处于其最高级的其他因子以及所有处于其均值的其他协变量。如果因子或协变量与模型中的其他预测变量交互作用，则几率比不仅取决于指定变量的变化，而且还取决于与之交互的变量的值。如果指定的协变量在模型中与其自身交互作用（例如 age*age），则几率比同时取决于协变量的变化和协变量的值。

复杂样本 Logistic 回归：保存

图片 10-7
“Logistic 回归：保存”对话框

保存变量。在这一组中，可以将模型预测类别和预测概率保存为活动数据集中的新变量。

将模型导出为 SPSS Statistics 数据。写入一个 IBM® SPSS® Statistics 格式的数据集，包含具有参数估计值、标准误、显著性值和自由度的参数相关性或协方差矩阵。矩阵文件中变量顺序如下。

- **RowType_。**取值（或值标签）为 COV（协方差）、CORR（相关性）、EST（参数估计）、SE（标准误）、SIG（显著性水平）和 DF（抽样设计自由度）。存在每个模型参数的 COV（或 CORR）行类型的单独个案，以及每个其他行类型的个案。
- **VarName_。**对于行类型 COV 或 CORR，取值为 P1、P2、...，对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串。对于其他行类型，单元格为空。
- **P1、P2、...**这些变量对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串，这些变量根据行类型取值。对于冗余参数，所有协方差设为零；相关性设为系统缺失值；所有参数估计值设为零；并且所有标准误、显著性水平和残差自由度设为系统缺失值。

注意：该文件不能立即用于在其他读取矩阵文件的过程中执行进一步分析，除非这些过程接受在此导出的所有行类型。

将模型导出为 XML。将参数估计值和参数协方差矩阵（如果选择）以 XML (PMML) 格式保存。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。

复杂样本 Logistic 回归：选项

图片 10-8
“Logistic 回归：选项”对话框

估值。在这一组中，可以控制模型估计中使用的各种条件。

- **最大迭代次数。**算法将执行的最大迭代次数。指定一个非负整数。
- **最大折半次数。**每次迭代时，步长都会减去因子 0.5，直到对数似然估计增加或者达到最大步骤对分。指定一个正整数。
- **根据参数估值更改限制迭代。**如果选择此项，算法将在参数估计值的绝对或相对变化小于指定值（必须为非负值）的迭代之后停止。
- **根据对数似然估计更改限制迭代。**如果选择此项，算法将在对数似然估计函数的绝对或相对变化小于指定值（必须为非负值）的迭代之后停止。
- **检查数据点的完整分隔。**如果选择此项，算法将执行检验以确保参数估计值具有唯一值。当过程可生成一个正确对每个个案进行分类的模型时，将发生分离。
- **显示迭代历史记录。**从第 0 次迭代（初始估计）开始，在每 n 次迭代时显示参数估计值和统计量。如果选择打印迭代历史记录，则无论 n 值为多少，将总是打印最后一次迭代。

用户缺失值。所有设计变量以及因变量和任何协变量都必须包含有效数据。从分析中删除含有任何这些变量的无效数据的个案。使用这些控制可以确定是否在层次、聚类、子体和因子变量中将用户缺失值视为有效。

置信区间。这是系数估值、取幂系数估值和几率比的置信区间度。指定大于等于 50 且小于 100 的值。

CSLOGISTIC 命令附加功能

使用命令语法语言还可以：

- 指定作用对比作用的线性组合或一个值的自定义检验（使用 CUSTOM 子命令）。
- 在计算因子和协变量的几率比（使用 ODDSRATIOS 子命令）时修正其他模型变量的值。
- 指定用于检查奇异性的容差值（使用 CRITERIA 子命令）。
- 为保存的变量创建用户指定名称（使用 SAVE 子命令）。
- 生成常规可估计函数表（使用 PRINT 子命令）。

请参阅命令语法参考以获取完整的语法信息。

章
11

复杂样本序数回归

“复杂样本序数回归”过程对通过复杂抽样方法抽取的样本的二元或序数因变量执行序数回归分析。您还可以请求对子体进行分析。

示例。议员在向立法院提交某项法案之前想了解公众是否支持该法案，以及对该法案的支持与选民人群统计信息有何关联。民意测验专家根据复杂抽样设计并实施了一些采访。使用复杂样本序数回归，可以根据选民人群统计信息将模型拟合到对方案的支持水平。

数据。因变量是序数变量。因子是分类变量。协变量是与因变量相关的定量变量。子体变量可以是字符串或数值，但应该是分类变量。

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

获取复杂样本序数回归

从菜单中选择：
分析 > **复杂抽样** > **序数回归...**

- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

图片 11-1
“Ordinal 回归”对话框

- 选择一个因变量。

根据需要，您可以：

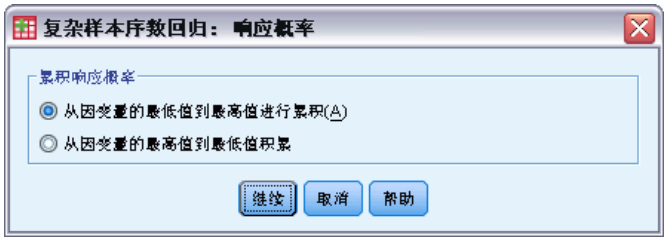
- 为因子和协变量选择适合您数据的变量。
- 指定用于定义子体的变量。尽管在整个数据集基础上对方差进行了正确估计，但仍只对所选的子体变量类别进行分析。
- 选择链接函数。

链接函数。链接函数是累积概率的转换形式，可用于模型估计。下表总结了五个可用的链接函数。

函数	形式	典型应用
Logit	$\log(\xi / (1-\xi))$	均匀分布类别
互补双对数	$\log(-\log(1-\xi))$	类别越高可能性越大
负双对数	$-\log(-\log(\xi))$	类别越低可能性越大
Probit	$\Phi^{-1}(\xi)$	潜在变量为正态分布
Cauchit（逆 Cauchy）	$\tan(\pi(\xi-0.5))$	潜在变量有许多个极值

复杂样本序数回归：响应概率

图片 11-2
“序数回归：响应概率”对话框



在“响应概率”对话框中，可以指定响应的累积概率是否随因变量值的增大或减小而增大，累积概率是属于因变量特定类别之前类别（包括该特定类别）的概率。

复杂样本序数回归：模型

图片 11-3
“序数回归模型”对话框

指定模型效应。缺省情况下，该过程使用主对话框中指定的因子和协变量构建主效应模型。此外，还可以构建包含交互效应和嵌套项的定制模型。

非嵌套项

对于选定因子和协变量：

交互。为所有选定变量创建最高级交互项。

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二阶交互。

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。

嵌套项

在此过程中，可为您的模型建立嵌套项。嵌套项有助于对其值不与另一个因子的水平交互作用的因子或协变量的效应进行建模。例如，杂货连锁店可能在不同商店位置迎合顾客的不同消费习惯。由于每位顾客只频繁光顾某一位置的商店，因此 Customer 效应可以说是**嵌套在** Store location 效应中。

此外，还可以包含交互效应，例如包含相同协变量的多项式项，或将多层嵌套添加到嵌套项。

限制。嵌套项有以下限制：

- 一次交互内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A*A 是无效的。
- 嵌套效应内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A(A) 是无效的。
- 效应不可嵌套在协变量中。因此，如果 A 是因子且 X 是协变量，则指定 A(X) 是无效的。

复杂样本序数回归：统计量

图片 11-4
“Ordinal 回归：统计量”对话框

模型拟合度。控制度量总体模型性能的统计量的显示。

- **伪 R 方。**在序数回归模型中，没有与线性回归的 R^2 对应的统计量。相反，却有多种试图模拟 R^2 统计量的属性的测量。
- **分类表。**按因变量的模型预测类别显示观察类别的制表交叉分类。

参数。使用此组可以控制与模型参数有关的统计量的显示。

- **估算。**显示系数的估计值。
- **取幂估值。**显示以系数估值为幂的自然对数的底数。当该估值对于统计检验有良好的属性时，取幂估值（即 exp(B)）更易于解释。
- **标准误。**显示每个系数估计值的标准误。
- **置信区间。**显示每个系数估计值的置信区间。在“选项”对话框中设置该区间的置信度。
- **T 检验。**显示每个系数估计值的 t 检验。每个检验的原假设是该系数的值为 0。
- **参数估值协方差。**显示模型系数的协方差矩阵的估计值。
- **参数估值的相关性。**显示模型系数的相关性矩阵的估计值。
- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**这是指定复杂设计的效果的测量值，以和标准误单位相当的单位表示，值与 1 相差越大表示效果越好。

平行线。在这一组中，可以请求与具有非平行线的模型关联的统计量，其中分别为各个响应类别（最后一个除外）单独拟合一个回归线。

- **Wald 检验。**生成回归参数对所有累积响应都相等的原假设检验。估计具有非平行线的模型并应用相等参数的 Wald 检验。
- **参数估计。**显示具有非平行线的模型的系数和标准误的估计值。
- **参数估值协方差。**显示具有非平行线的模型的系数的协方差矩阵估计值。

模型变量的摘要统计。显示有关因变量、协变量和因子的摘要信息。

样本设计信息。显示有关样本的摘要信息，包括未加权的计数和总体大小。

复杂样本假设检验

图片 11-5
“假设检验”对话框

检验统计。在这一组中，可以选择用于检验假设的统计类型。可以在 F、调整的 F、卡方和调整的卡方之间选择。

样本自由度。在这一组中，可以控制用于计算所有检验统计量的 p 值的抽样设计自由度。如果基于抽样设计，该值为抽样第一阶段的主抽样单元数和层数之差。或者，也可以通过指定一个正整数设置定制自由度。

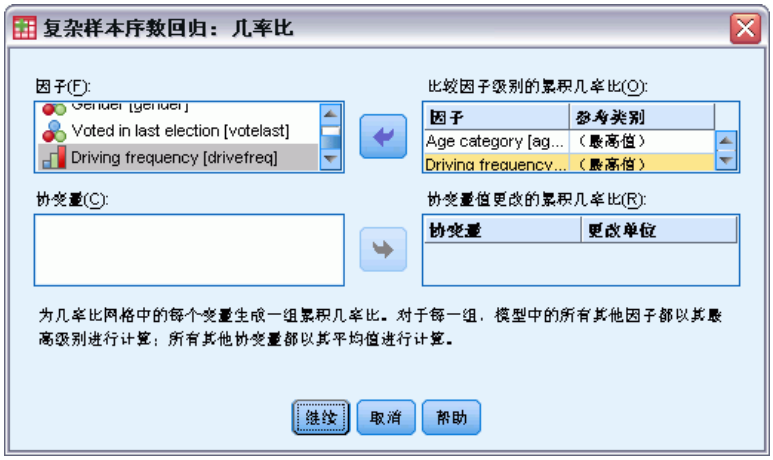
调整的多重比较。在执行包含多重比较的假设检验时，总体显著性水平可从所包含的比较的显著性水平进行调节。使用此组可以选择调节方法。

- **显著性最低的差异。**此方法并不控制拒绝某些线性对比不同于原假设值这一假设的总体概率。
- **连续 Sidak.** 这是按顺序逐步降低的拒绝 Sidak 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。

- **连续 Bonferroni.** 这是按顺序逐步降低的拒绝 Bonferroni 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。
- **Sidak.** 此方法提供比 Bonferroni 方法更严密的界限。
- **Bonferroni.** 此方法针对检验多个对比这一事实调整观测的显著性水平。

复杂样本序数回归：几率比

图片 11-6
“Ordinal 回归：几率比”对话框



使用“几率比”对话框，可以控制指定的因子和协变量的模型估计累积几率比的显示。此功能仅用于使用 Logit 关联函数的模型。计算因变量所有类别（最后一个除外）的单个累积几率比；比例几率模型假设它们都相等。

因子。对于所选的每个因子，都显示该因子处于每个类别的累积几率与处于指定参考类别的几率之比。

协变量。对于所选的每个协变量，都显示协变量均值加上指定变化单位的累积几率与处于均值的几率之比。

在计算因子或协变量的几率比时，该过程将修正所有处于其最高级的其他因子以及所有处于其均值的其他协变量。如果因子或协变量与模型中的其他预测变量交互作用，则几率比不仅取决于指定变量的变化，而且还取决于与之交互的变量的值。如果指定的协变量在模型中与其自身交互作用（例如 age*age），则几率比同时取决于协变量的变化和协变量的值。

复杂样本序数回归：保存

图片 11-7
“序数回归保存”对话框

保存变量。在这一组中，可以将模型预测类别、预测类别的概率、观察类别的概率、累积概率和预测概率保存为活动数据集中的新变量。

将模型导出为 SPSS Statistics 数据。写入一个 IBM® SPSS® Statistics 格式的数据集，包含具有参数估计值、标准误、显著性值和自由度的参数相关性或协方差矩阵。矩阵文件中变量顺序如下。

- **RowType_。**取值（或值标签）为 COV（协方差）、CORR（相关性）、EST（参数估计）、SE（标准误）、SIG（显著性水平）和 DF（抽样设计自由度）。存在每个模型参数的 COV（或 CORR）行类型的单独个案，以及每个其他行类型的个案。
- **VarName_。**对于行类型 COV 或 CORR，取值为 P1、P2、...，对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串。对于其他行类型，单元格为空。
- **P1、P2、...**这些变量对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串，这些变量根据行类型取值。对于冗余参数，所有协方差设为零；相关性设为系统缺失值；所有参数估计值设为零；并且所有标准误、显著性水平和残差自由度设为系统缺失值。

注意：该文件不能立即用于在其他读取矩阵文件的过程中执行进一步分析，除非这些过程接受在此导出的所有行类型。

将模型导出为 XML。将参数估计值和参数协方差矩阵（如果选择）以 XML (PMML) 格式保存。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。

复杂样本序数回归：选项

图片 11-8
Ordinal 回归：选项对话框

估计方法。可以选择参数估计方法；其中包括：Newton-Raphson、Fisher 评分方法以及先执行 Fisher 评分迭代再切换为 Newton-Raphson 方法的混合方法。如果在混合方法的 Fisher 评分方法阶段期间，在达到 Fisher 迭代的最大次数之前实现了收敛，则算法将继续执行 Newton-Raphson 方法。

估值。在这一组中，可以控制模型估计中使用的各种条件。

- **最大迭代次数。**算法将执行的最大迭代次数。指定一个非负整数。
- **最大折半次数。**每次迭代时，步长都会减去因子 0.5，直到对数似然估计增加或者达到最大步骤对分。指定一个正整数。
- **根据参数估值更改限制迭代。**如果选择此项，算法将在参数估计值的绝对或相对变化小于指定值（必须为非负值）的迭代之后停止。
- **根据对数似然估计更改限制迭代。**如果选择此项，算法将在对数似然估计函数的绝对或相对变化小于指定值（必须为非负值）的迭代之后停止。
- **检查数据点的完整分隔。**如果选择此项，算法将执行检验以确保参数估计值具有唯一值。当过程可生成一个正确对每个个案进行分类的模型时，将发生分离。
- **显示迭代历史记录。**从第 0 次迭代（初始估计）开始，在每 n 次迭代时显示参数估计值和统计量。如果选择打印迭代历史记录，则无论 n 值为多少，将总是打印最后一次迭代。

用户缺失值。刻度设计变量以及因变量和任何协变量都应包含有效数据。从分析中删除含有任何这些变量的无效数据的个案。使用这些控制可以确定是否在层次、聚类、子体和因子变量中将用户缺失值视为有效。

置信区间。这是系数估值、取幂系数估值和几率比的置信区间度。指定大于等于 50 且小于 100 的值。

CSORDINAL 命令附加功能

使用命令语法语言还可以：

- 指定作用对比作用的线性组合或一个值的自定义检验（使用 CUSTOM 子命令）。
- 在计算因子和协变量的累积几率比（使用 ODDSRATIOS 子命令）时，修正其他不处于其均值的模型变量值。
- 在请求几率比（使用 ODDSRATIOS 子命令）时，使用未标注值作为因子的自定义参考类别。
- 指定用于检查奇异性的容差值（使用 CRITERIA 子命令）。
- 生成常规可估计函数表（使用 PRINT 子命令）。
- 保存 25 个以上的概率变量（使用 SAVE 子命令）。

请参阅命令语法参考以获取完整的语法信息。

章
12

复杂样本 Cox 回归

复杂样本 Cox 回归过程对由复杂取样方法抽取的样本进行生存分析。您还可以请求对子体进行分析。

示例。 政府执法机构关心其管辖区域内的屡犯率。测量屡犯率的方法之一就是罪犯第二次被捕的时间。机构希望利用 Cox 回归对时间建模以进行再次抓捕，但又担心成比例的风险假定在跨越年龄类别时失效。

医疗研究者正在对结束肌肉萎缩后中风症状复原计划后的患者存活时间进行调查。自从患者病史如所记录的显著非死亡事件的发生和次数而变化以来就存在每位主体具有多项个案的潜在性。就观察存活时间被复元长度“夸大”的意义而言，样本同样为左侧截短，因为当肌肉萎缩中风的风险攻击开始产生时，只有在复元计划后存活下来的患者存在于样本中。

生存时间。 将 Cox 回归应用于存活时间分析的过程 — 为事件发生前的时间长度。有两种方式特定存活时间，取决于区间的开始时间：

- **Time=0。** 通常情况下，将在每个主体的区间开始具有完整信息并仅具有包含结束时间（或从“日期 & 时间”变量中创建一个带有结束时间的单一变量；参见如下）的变量。
- **依对象变化。** 这在您具有**左侧截短**，又称**延迟条目**时很恰当；例如，您正在对结束中风后症状复原计划后的患者存活时间进行分析，您可能会考虑到他们在中风时开始产生的风险攻击。但是，若您的样本只包含复元计划中存活的患者，就观察的存活时间被复元长度“夸大”的意义而言，则您的样本为左侧截短。可以通过将他们结束复元的时间指定为进入研究的时间对此进行说明。

日期 & 时间变量。“日期 & 时间”变量不能被用于直接界定区间的起始点与结束点；您若具有“日期 & 时间”就应该用其创建包含存活时间的变量。若无左侧截短，则根据进入研究的日期与观察日期间的差异仅创建包含结束时间的变量。若无左侧截短，则根据开始研究的日期与进入日期间的差异创建包含开始时间的变量并根据开始研究的日期与观察日期间的差异创建包含结束时间的变量。

事件状态。 您需要记录了主体在区间中是否经历过被观察事件的变量。事件并非为其发生的主体为右侧已审查。

对象识别。 您可以通过拆分跨多项个案的单一主体观察轻松合并分段恒定、依时预测器。例如，若您正在分析中风后患者存活时间，代表其医疗历史的变量应该是有用的预测器。随着时间变化，他们可能会经历改变其医疗历史的重要医疗事件。以下表格说明如何构建这种数据集：Patient ID 为主体标识，End time 界定被观察的区间，

Status 记录重要医疗事件，且 Prior history of heart attack 与 Prior history of hemorrhaging 为分段恒定、依时预测器。

Patient ID	End time	状态	Prior history of heart attack	Prior history of hemorrhaging
1	5	心脏病	No	No
1	7	出血病	Yes	No
1	8	死亡	Yes	Yes
2	24	死亡	No	No
3	8	心脏病	No	No
3	15	死亡	Yes	No

假设。数据文件中的个案代表来自复杂设计的一个样本，该样本应根据在“复杂样本计划”对话框中所选文件内的指定项进行分析。

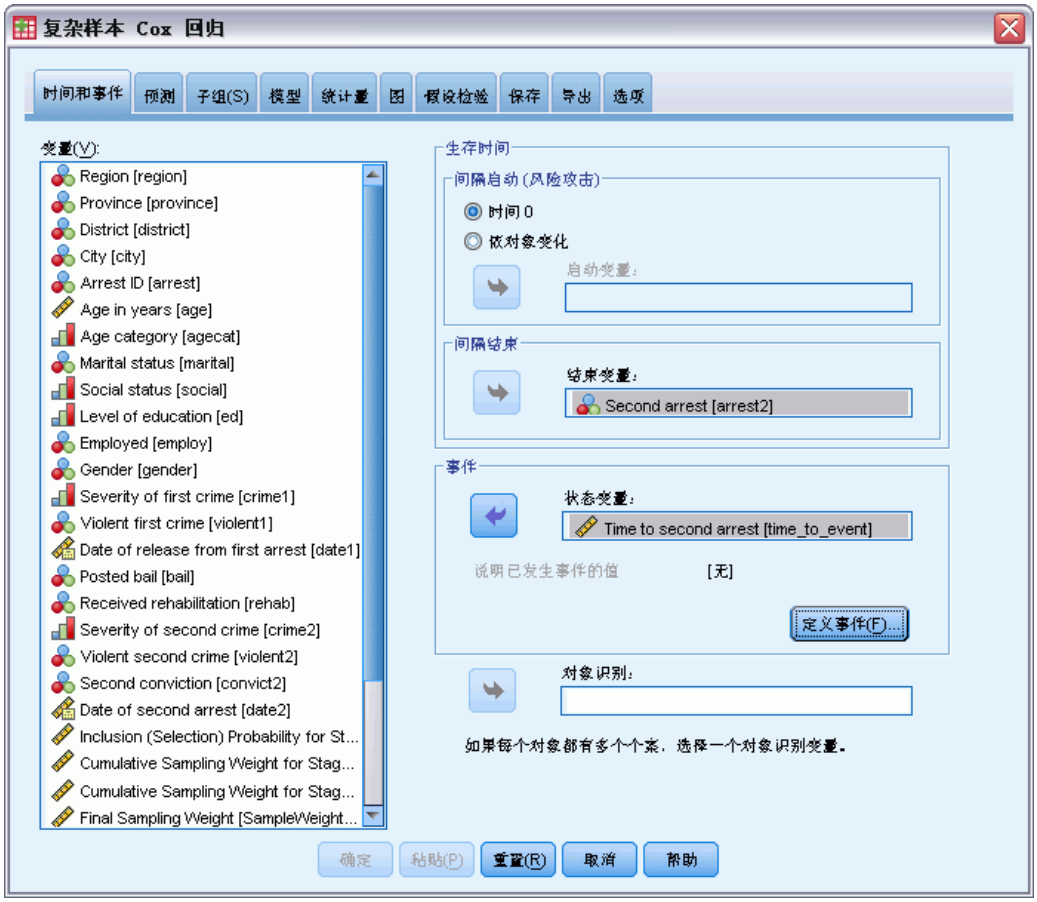
通常情况下，Cox 回归 模型假定成比例的风险 — 即从一个个案到另一个个案间的风险比不应随时间变化。若此假设不成立，可能需要将依时预测器添加至模型。

Kaplan-Meier 分析。 您若不选择任何预测器（或不向模型中输入任何已选预测器）而选择在“选项”选项卡中利用产品限制方法计算基线存活曲线，则该过程将进行 Kaplan-Meier 型存活分析。

获取复杂样本 Cox 回归

- 从菜单中选择：
  分析 > 复杂抽样 > Cox 回归...
- 选择计划文件。根据需要，选择客户加入概率文件。
- 单击**继续**。

图片 12-1
“Cox 回归”对话框，“时间与事件”选项卡

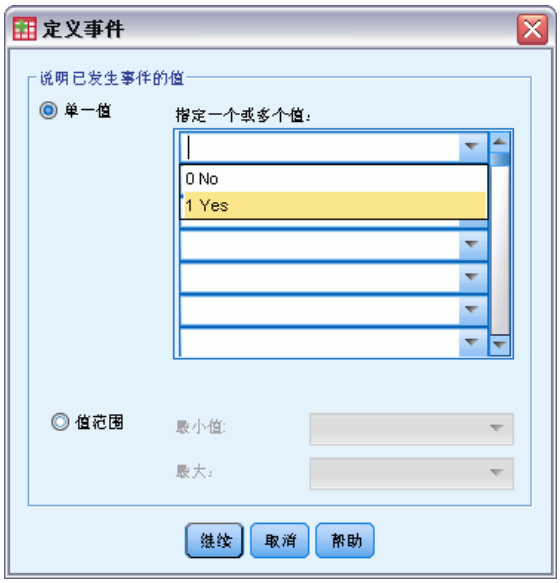


- 通过选择进入和退出研究的时间指定存活时间。
- 选择事件状态变量。
- 单击界定事件并界定至少一个事件值。

  根据需要，可以选择主体标识。

界定事件

图片 12-2
“定义事件”对话框

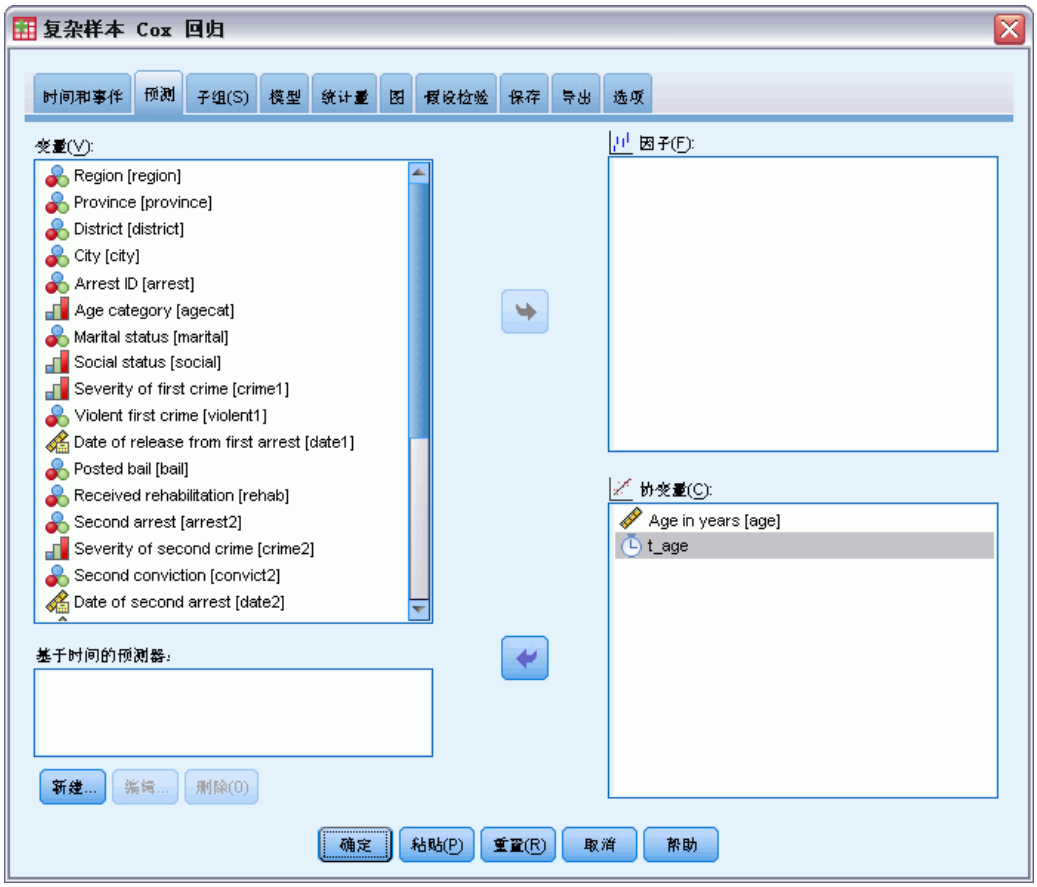


指定能指明终结事件已发生的值。

- **单一值。** 通过将其输入网格或从带有界定值标签的列表中将其选中以指定一个或多个值。
- **值范围。** 通过输入最小值和最大值或从带有界定值标签的列表中选择值以指定一系列值。

预测器

图片 12-3
“Cox 回归”对话框，“预测器”选项卡



“预测器”选项卡可以指定用于建立模型效应的因子及协变量。

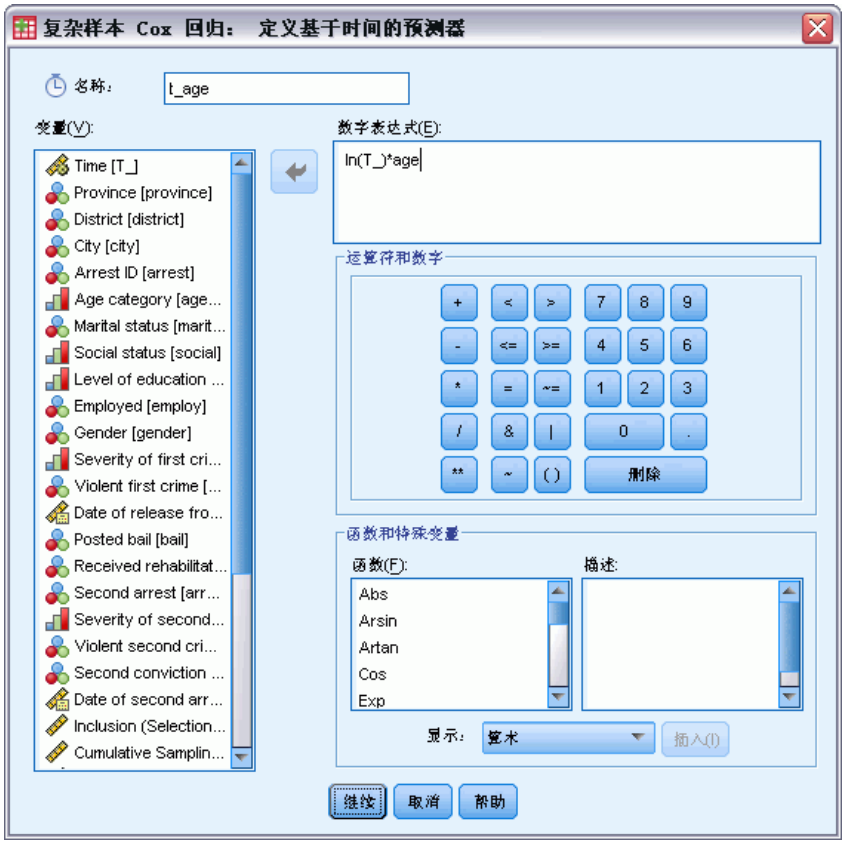
因子。因子是分类预测变量，可以是数值或字符串。

协变量。协变量为刻度预测变量，必须为数值。

基于时间的预测器。 存在某些成比例的风险假设不成立的情况。也就是说，风险比率随时间变化；在不同的时间点一个（或多个）预测器的值会不同。在此情况下，需要指定依时预测器。有关详细信息，请参阅第 72 页码中的界定依时预测器．依时预测器可作为因子或协变量选择。

界定依时预测器

图片 12-4
“Cox 回归界定依时预测器”对话框



“界定依时预测器”对话框可以创建依据内嵌时间变量的预测器，T_。您可以使用此变量通过两种常用方法定义依时协变量：

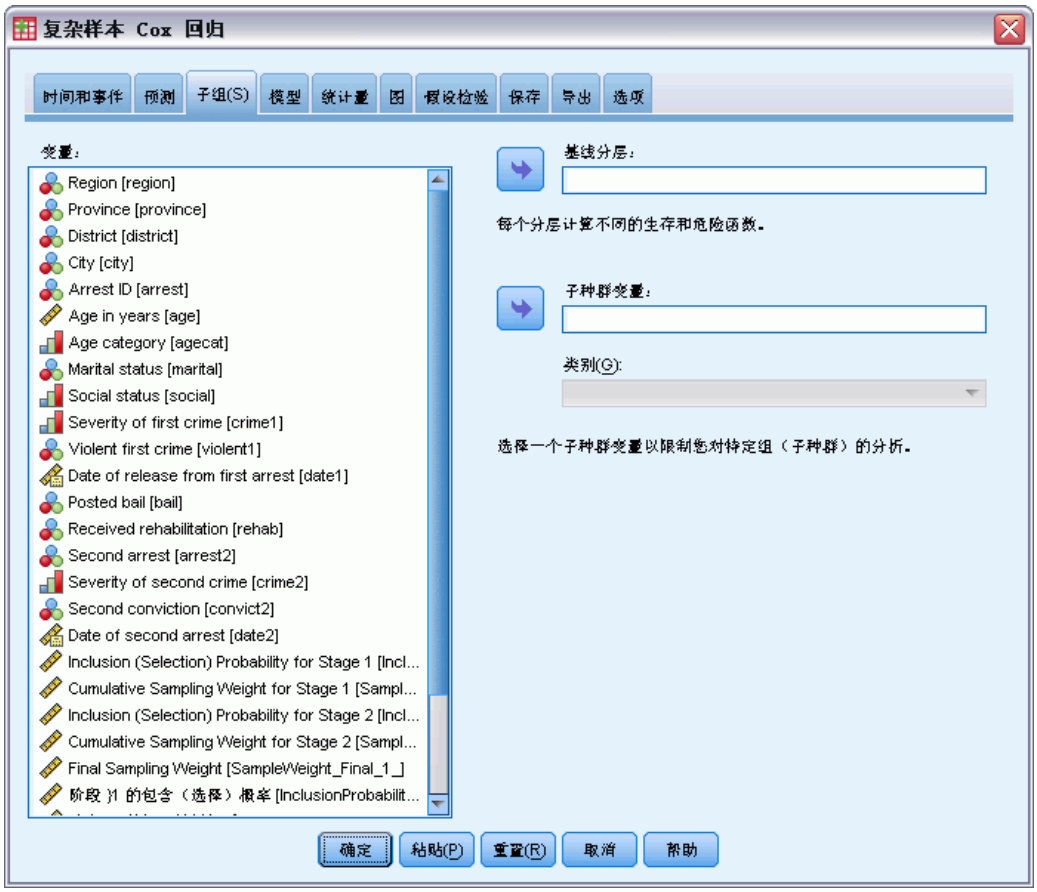
- 若想对允许非成比例风险的扩展Cox 回归模型进行估计，可通过将依时预测器界定为问题中的时间变量 T_ 与协变量功能来实现。一个常见的例子就是简单地将时间变量和预测器相乘，不过也可以指定较为复杂的函数。
- 有些变量在不同的时间段内可能具有不同的值，但其值与时间并不具有系统相关性。在这样的情况下，您需要定义一个**分段依时预测器**，这可以通过使用逻辑表达式完成。逻辑表达式使用值 1 表示“true”，使用值 0 表示“false”。通过使用一系列逻辑表达式，您就可以使用一组度量创建依时预测器。例如，如果您在一个为期四周的研究中每周测量一次血压（使用 BP1 到 BP4 标识），则可以将依时预测定义为 (T_ < 1) * BP1 + (T_ >= 1 & T_ < 2) * BP2 + (T_ >= 2 & T_ < 3) * BP3 + (T_ >= 3 & T_ < 4) * BP4。注意，对于任何给定的个案，括号中都正好有一个项等于 1，其余项都等于 0。换言之，此函数意味着如果时间小于一周则使用 BP1；如果时间大于一周但小于两周则使用 BP2；依此类推。

  注意： 若分段的依时预测器如以上给出的血压示例一样与段恒定，则更容易通过拆分跨越多项个案的主体指定分段恒定、依时预测器。参见 复杂样本 Cox 回归第 67 页码 中“主体标识”中的讨论获取更多信息。

在“界定依时预测器”对话框中，您可以使用函数构建控件构建依时协变量的表达式，或者可以在“数字表达式”文本区域中直接输入表达式。注意，字符串常数必须包含在引号或单引号中，数字常数必须以美式格式键入，并使用句点作为小数定界符。结果变量被赋予指定的名称并应作为因子或协变量包含在“预测器”选项卡中。

子组

图片 12-5
“Cox 回归”，“子组”选项卡

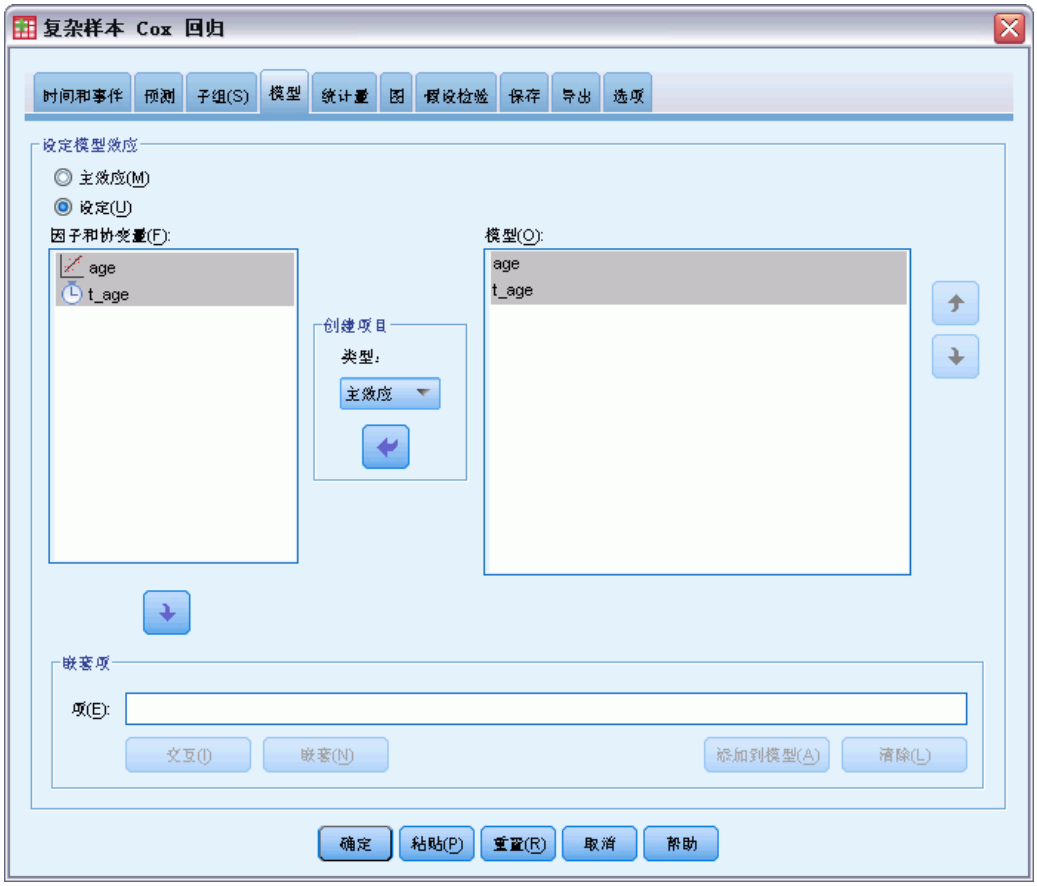


基线分层。 为每个变量值计算单独基线风险和生存功能的同时，跨层估计一个单一模型系数集。

子体变量。 指定用于定义子体的变量。仅对子体变量的所选类别执行该分析。

模型

图片 12-6
“Cox 回归”，“模型”选项卡



指定模型效应。默认情况下，该过程使用主对话框中指定的因子和协变量构建主效应模型。此外，还可以构建包含交互效应和嵌套项的自定义模型。

非嵌套项

对于选定因子和协变量：

交互。为所有选定变量创建最高级交互项。

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二阶交互。

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。

嵌套项

在此过程中，可为您的模型建立嵌套项。嵌套项有助于对其值不与另一个因子的水平交互作用的因子或协变量的效应进行建模。例如，杂货连锁店可能在不同商店位置迎合顾客的不同消费习惯。由于每位顾客只频繁光顾某一位置的商店，因此 Customer 效应可以说是**嵌套在** Store location 效应中。

此外，还可以包含交互效应，例如包含相同协变量的多项式项，或将多层嵌套添加到嵌套项。

限制。嵌套项有以下限制：

- 一次交互内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A*A 是无效的。
- 嵌套效应内的所有因子必须是唯一的。因此，如果 A 是因子，则指定 A(A) 是无效的。
- 效应不可嵌套在协变量中。因此，如果 A 是因子且 X 是协变量，则指定 A(X) 是无效的。

统计量

图片 12-7
“Cox 回归”对话框，“统计量”选项卡

样本设计信息。显示有关样本的摘要信息，包括未加权的计数和总体大小。

事件与检查摘要。 显示关于已审查个案数量与比例的摘要信息。

事件时间中的风险设置。 显示每个基线分层中每个事件时间的事件数量和带风险数量。

参数。使用此组可以控制与模型参数有关的统计量的显示。

- **估算。**显示系数的估计值。
- **取幂估值。**显示以系数估值为幂的自然对数的底数。当该估值对于统计检验有良好的属性时，取幂估值（即 exp(B)）更易于解释。
- **标准误。**显示每个系数估计值的标准误。
- **置信区间。**显示每个系数估计值的置信区间。在“选项”对话框中设置该区间的置信度。
- **t 检验。**显示每个系数估计值的 t 检验。每个检验的原假设是该系数的值为 0。
- **参数估值协方差。**显示模型系数的协方差矩阵的估计值。
- **参数估值的相关性。**显示模型系数的相关性矩阵的估计值。
- **设计效应。**估计值的方差与通过假设样本为简单随机样本所获得的方差的比率。这是指定复杂设计的效果测量，该值与 1 相差越大，表示效果越大。
- **设计效应的平方根。**是指定复杂设计的效果的测量值，值与 1 相差越大表示效果越好。

模型假设。 此组可生成对成比例风险假设的检测。该检测将拟合模型与包含依时预测器 x*_TF 的备用模型做对比，每个预测器 x 都具有 _TF 特定时间函数。

- **时间函数。** 指定备用模型 _TF 的形式。对于**恒等**函数，_TF=T_。对于**对数**函数，_TF=T_。对于 **Kaplan-Meier**，_TF=1−S_{KM}(T_)，其中 S_{KM}(.) 生存函数的 Kaplan-Meier 估计。对于**秩**，_TF 在被观察的结束时间中是 T_ 的秩次。
- **其他模型的参数估计值。** 在备用模型中显示每个参数的估计、标准误和置信区间。
- **其他模型的协变量矩阵。** 在备用模型中显示参数间估计的协方差矩阵。

基线生存和累积危险函数。 连同其标准误一同显示基线生存函数与基线累积风险。

注意：若“预测器”选项卡中界定的依时预测器包含在模型中，则此选项不可用。

图

图片 12-8
“Cox 回归”，“图”选项卡

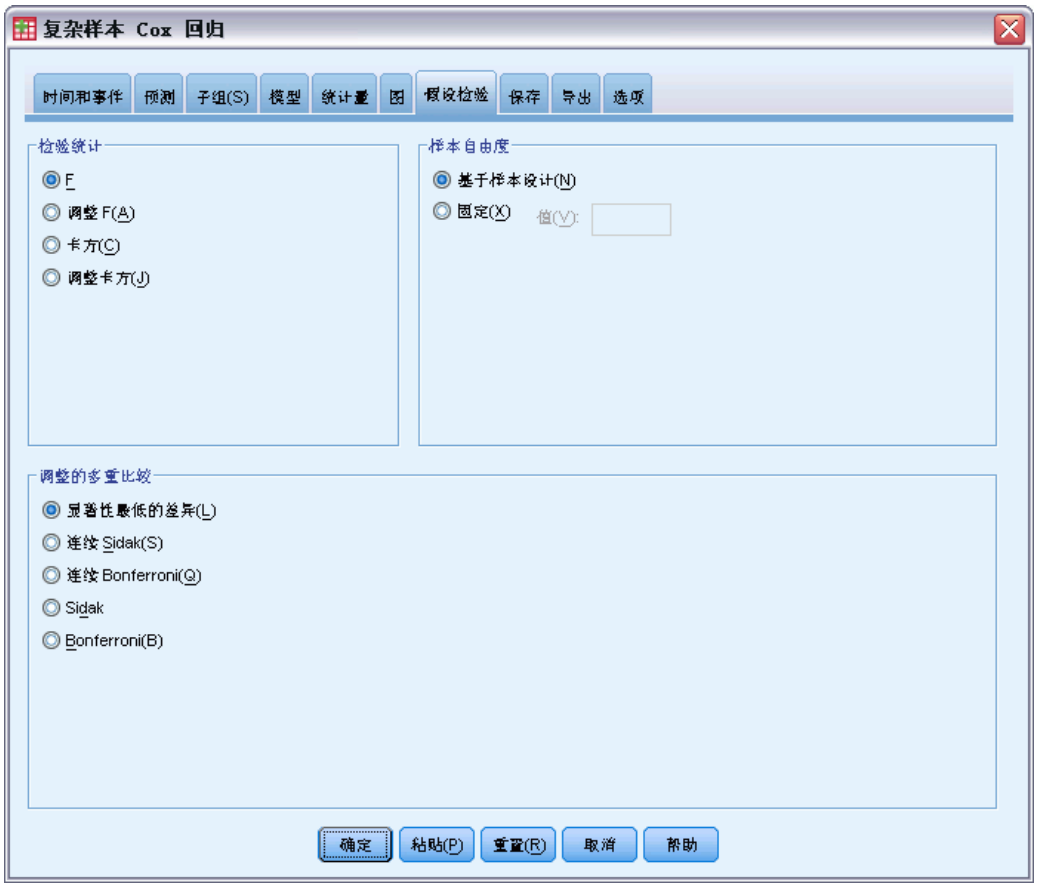
“图”选项卡可以申请风险函数图、生存函数图、对数负对数生存函数图以及 1 减生存函数图。还可以选择根据特定函数绘制置信区间图；在“选项”选项卡中设置置信水平。

预测器模式。可以指定预测器值模式用于要求图与“导出”选项卡中的导出生存文件。注意，若“预测器”选项卡中界定的依时预测器包含在模型中，则此选项不可用。

- **图表因子在。** 默认情况下，每个因子以其最高水平进行评估。若有需要请输入或选择不同水平。或者，还可以选择通过为该因子选择复选框为单一因子的每个水平绘制分离线。
- **图表协变量在。** 每个协变量以其均值进行评估。若有需要请输入或选择不同值。

假设检验

图片 12-9
“Cox 回归”，“假设检验”选项卡



检验统计。在这一组中，可以选择用于检验假设的统计类型。可以在 F、调整的 F、卡方和调整的卡方之间选择。

样本自由度。在这一组中，可以控制用于计算所有检验统计量的 p 值的抽样设计自由度。如果基于抽样设计，该值为抽样第一阶段的主抽样单元数和层数之差。或者，也可以通过指定一个正整数设置自定义自由度。

调整的多重比较。在执行包含多重比较的假设检验时，总体显著性水平可从所包含的比较的显著性水平进行调节。使用此组可以选择调节方法。

- **显著性最低的差异。**此方法并不控制拒绝某些线性对比不同于原假设值这一假设的总体概率。
- **连续 Sidak.** 这是按顺序逐步降低的拒绝 Sidak 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。
- **连续 Bonferroni.** 这是按顺序逐步降低的拒绝 Bonferroni 过程，在拒绝个别假设方面不保守，但维持相同的总体显著性水平。
- **Sidak.** 此方法提供比 Bonferroni 方法更严密的界限。
- **Bonferroni.** 此方法针对检验多个对比这一事实调整观测的显著性水平。

保存

图片 12-10
“Cox 回归”，“保存”选项卡

保存变量。 此组可以将相关模型变量保存至活动数据集以备将来用于诊断和结果报告。注意，当依时预测器包含在模型中时，全部不可用。

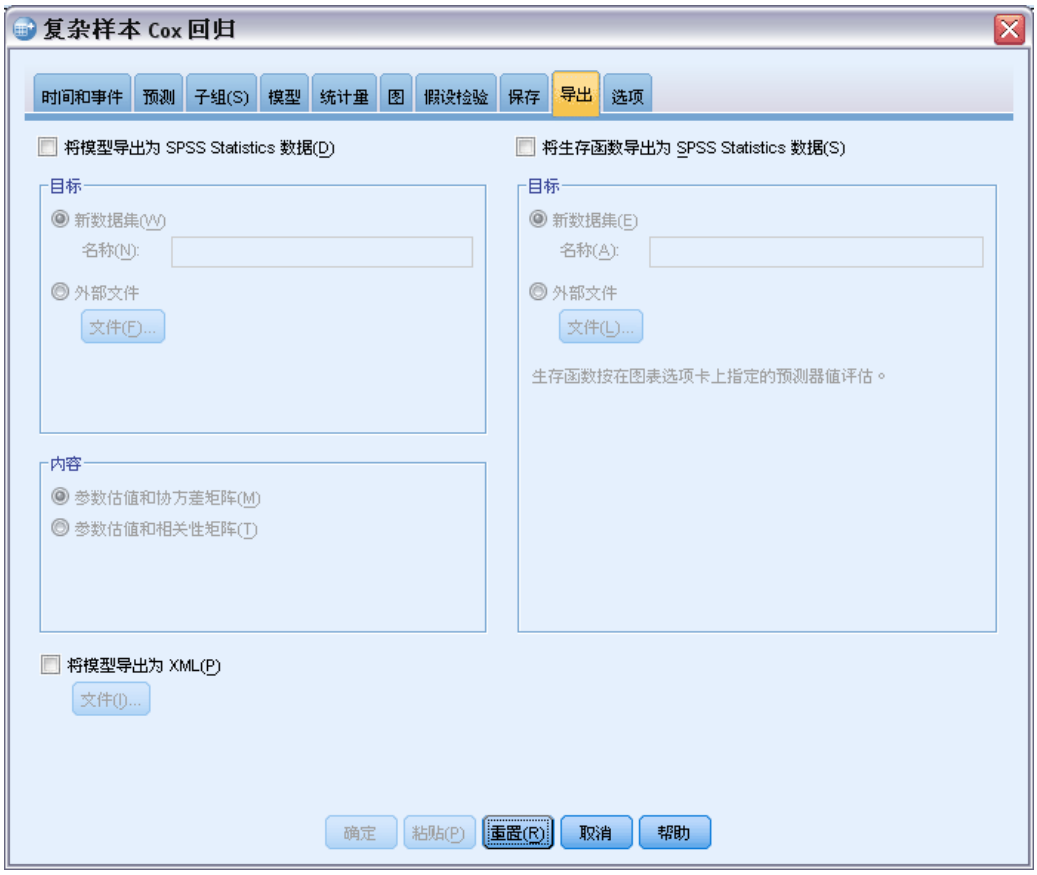
- **生存函数。** 以被观察时间和每个个案的预测器值保存生存概率（生存函数值）。
- **生存函数的置信区间下限。** 以被观察时间和每个个案的预测器值保存生存函数的置信区间下限。
- **生存函数的置信区间上限。** 以被观察时间和每个个案的预测器值保存生存函数的置信区间上限。
- **累积危险函数。** 以被观察时间和每个个案的预测器值保存累积风险，或 –在（生存函数）中。
- **累积危险函数的置信区间下限。** 以被观察时间和每个个案的预测器值保存累积危险函数的置信区间下限。
- **累积危险函数的置信区间上限。** 以被观察时间和每个个案的预测器值保存累积危险函数的置信区间上限。
- **线性预测器的预测值。** 保存参考值的线性组合，更正预测器时间回归系数。线性预测器是风险函数占基线风险的比例。在成比例风险模型下，此值为时间恒定值。

- **Schoenfeld 残差。** 对于模型中的每个为审查个案和非冗余参数，Schoenfeld 残差是与模型参数相关的被观察预测器值与以被观察时间设置的带风险个案预期预测器值之间的差。Schoenfeld 残差可用于帮助使用成比例风险假设；例如，对于预测器 x，若成比例风险成立，依时预测器 x*ln(T_) 与时间的 Schoenfeld 残差图应该在 0 处显示一条水平线。为模型中的每个非冗余参数保存单独变量。Schoenfeld 残差只为为审查个案计算。
- **Martingale 残差。** 对于每个个案，martingale 残差是被观察审查（审查为 0，未审查为 1）与观察时间内的事件预期值之间的差。
- **偏差残差。** 偏差残差是 martingale 残差“调整”后关于 0 更加对称的表现。偏差残差对于预测器的图不应揭示任何模式。
- **Cox-Snell 残差。** 对于每个个案，Cox-Snell 残差是观察时间内的事件预期值或被观察审查减去 martingale 残差。
- **Score 残差。** 对于每个个案和模型中的冗余参数，score 残差是第一个伪似然导数个案的贡献。为模型中的每个非冗余参数保存单独变量。
- **DFBeta 残差。** 对于每个个案和模型中的冗余参数，当个案从模型中被删除时，DFBeta 残差近似于参数估计值的变化。带有大型 DFBeta 残差的个案可能会对分析施加过度影响。为模型中的每个非冗余参数保存单独变量。
- **汇总残差。** 当多项个案代表一个单一主体时，主体的汇总残差只是全部个案中属于同一主体的相应个案残差的总和。对于 Schoenfeld’s 残差，总汇版本与非总汇版本相同，因为 Schoenfeld’s 残差只为未审查个案界定。这些残差只有当主体标识在“时间与事件”选项卡中被指定时方可使用。

保存的变量名称。 自动名称生成以确保您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。

导出

图片 12-11
“Cox 回归”，“导出”选项卡



将模型导出为 SPSS Statistics 数据。写入一个 IBM® SPSS® Statistics 格式的数据集，包含具有参数估计值、标准误、显著性值和自由度的参数相关性或协方差矩阵。矩阵文件中变量顺序如下。

- **RowType_。**取值（或值标签）为 COV（协方差）、CORR（相关性）、EST（参数估计）、SE（标准误）、SIG（显著性水平）和 DF（抽样设计自由度）。存在每个模型参数的 COV（或 CORR）行类型的单独个案，以及每个其他行类型的个案。
- **VarName_。**对于行类型 COV 或 CORR，取值为 P1、P2、...，对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串。对于其他行类型，单元格为空。
- **P1、P2、...**这些变量对应于所有模型参数的有序列表，值标签对应于在参数估计值表中显示的参数字符串，这些变量根据行类型取值。对于冗余参数，所有协方差设为零；相关性设为系统缺失值；所有参数估计值设为零；并且所有标准误、显著性水平和残差自由度设为系统缺失值。

注意：该文件不能立即用于在其他读取矩阵文件的过程中执行进一步分析，除非这些过程接受在此导出的所有行类型。

将生存函数导出为 SPSS Statistics 数据。为每个失败或事件时间编写一个包含生存函数的 SPSS Statistics 格式数据集；生存函数标准误；生存函数置信区间上下限；以及累积风险函数，并以基线和在“图”选项卡中指定的预测器模式对其进行评估。矩阵文件中变量顺序如下。

- **基线分层变量。** 为每个分层变量值生成单独的生存表格。
- **生存时间变量。** 事件时间；为每个独有的事件时间创建一个单独个案。
- **Sur_0, LCL_Sur_0, UCL_Sur_0。** 基线生存函数与其置信区间的上下限。
- **Sur_R, LCL_Sur_R, UCL_Sur_R。** 以“参考”模式评估的生存函数（参见输出中的模式值表格）与其置信区间的上下限。
- **Sur_#.#, LCL_Sur_#.#, UCL_Sur_#.#, …** 以在“图”选项卡中指定的每种预测器模式评估的生存函数与其置信区间的上下限。参见输出中的模式值表格与带有数字 #.# 的模式匹配。
- **Haz_0, LCL_Haz_0, UCL_Haz_0。** 基线累积风险函数与其置信区间的上下限。
- **Haz_R, LCL_Haz_R, UCL_Haz_R。** 以“参考”模式评估的累积风险（参见输出中的模式值表格）与其置信区间的上下限。
- **Haz_#.#, LCL_Haz_#.#, UCL_Haz_#.#, …** 以在“图”选项卡中指定的每种预测器模式评估的累积风险函数与其置信区间的上下限。参见输出中的模式值表格与带有数字 #.# 的模式匹配。

将模型导出为 XML。保存预测生存函数所需的全部信息，包括 XML（PMML）格式的参数评估与基线生存函数。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。

选项

图片 12-12
“Cox 回归”，“选项”选项卡

估值。 这些控件为回归系数的评估指定准则。

- **最大迭代次数。**算法将执行的最大迭代次数。指定一个非负整数。
- **最大折半次数。**每次迭代时，步长都会减去因子 0.5，直到对数似然估计增加或者达到最大步骤对分。指定一个正整数。
- **根据参数估值更改限制迭代。**如果选择此项，算法将在参数估计值的绝对或相对更改小于指定值（必须为正值）的迭代之后停止。
- **根据对数似然估计更改限制迭代。**如果选择此项，算法将在对数似然估计函数的绝对或相对更改小于指定值（必须为正值）的迭代之后停止。
- **显示迭代历史记录。**显示参数估计值和伪对数似然估计的迭代历史记录，并打印对参数评估变化和伪对数似然的最后一次评估。迭代历史记录表打印从第 0 次迭代（初始估计值）开始的每 n 次迭代，其中 n 代表增量值。如果请求迭代历史记录，那么无论 n 的值是多少，都会显示最后一次迭代。
- **参数估算的断开连接方法。** 当出现绑定的被观察失败时间时，此种方法之一将用于断开绑定。Efron 方法为更加大量的计算。

生存函数。 这些控件为涉及生存函数的计算指定准则。

- **基线生存函数的估算方法。**Breslow（或 Nelson-Aalan 或经验的）方法通过非减少步骤函数与被观察失败时间共同评估基线累积风险，然后通过关系 survival=exp(−cumulative hazard）计算基线生存。Efron 方法为更加大量的计算并当无绑定时简化为 Breslow 方法。**产品限制**方法通过非增加的正确连续函数评估基线生存；当模型中不存在任何预测器时，此方法简化为 Kaplan-Meier 评估。
- **生存函数的置信区间。** 可通过以下三种方式对置信区间进行计算：在初始单位中、通过对数转换，或对数负对数转换。只有对数负对数转换能保证置信区间限介于 0 和 1 之间，但对数转换被普遍认为是性能“最好”的方法。

用户缺失值。要在分析中包含个案，所有变量必须具有有效值。这些控件可以决定用户缺失值在类别模型（包括因子、事件、层次和子体变量）和取样设计变量中是否有效。

置信区间 (%)。 这是用于系数估计值、取幂系数估计值以及累积风险函数估计值的置信区间水平。指定大于等于 0 且小于 100 的值。

CSCOXREG 命令的附加功能

使用命令语言还可以：

- 执行自定义的假设检验（使用 CUSTOM 子命令和 /PRINT LMATRIX）。
- 容差规范（使用 /CRITERIA SINGULAR）。
- 一般可估计函数表格（使用 /PRINT GEF）。
- 多种预测器模式（使用多个 PATTERN 子集）。
- 当根名称被指定时（使用 SAVE 子命令），将已保存变量的数量最大化。对话框接受 CSCOXREG 25 个变量的默认。

请参阅命令语法参考以获取完整的语法信息。

部分 II:
示例

章
13

复杂样本抽样向导

该抽样向导将指导您完成创建、修改或执行抽样计划文件的步骤。在使用向导之前，应构思好定义明确的目标总体、抽样单元列表和适当的样本设计。

从完整抽样框架获取样本

一个州政府机构负责确保各县资产税的公平性。税是根据资产评估值确定的，因此，该机构希望调查各县的资产样本，以确保每个县的记录都是最新的。但是，获取当前评估的资源是有限的，因此合理使用可用资源就很重要。该机构决定使用复杂抽样方法来选择资产样本。

资产列表收集在 property_assess_cs.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.使用复杂样本抽样向导选择样本。

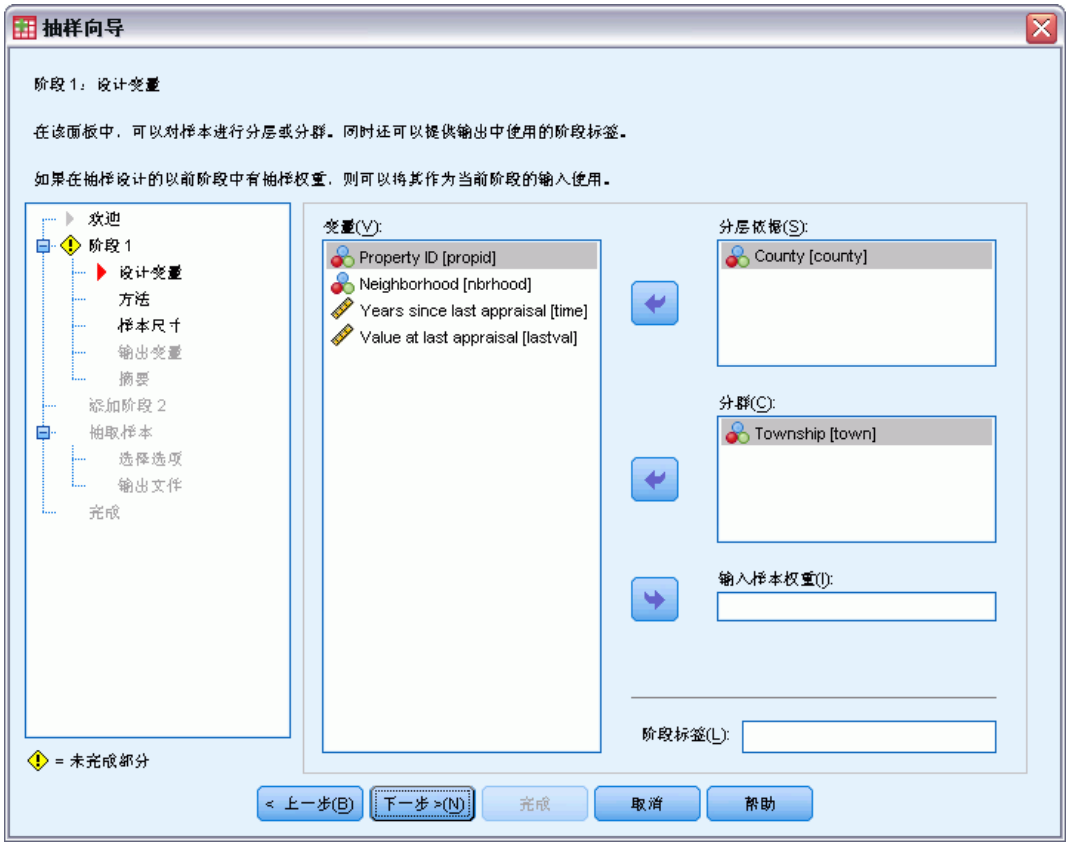
使用向导

- 要运行复杂样本抽样向导，请从菜单中选择：
  分析 > **复杂抽样** > **选择样本...**

图片 13-1
抽样向导，“欢迎”步骤

- 选择**设计样本**，浏览到要保存文件的位置，并输入 `property_assess.csplan` 作为计划文件的名称。
- 单击**下一步**。

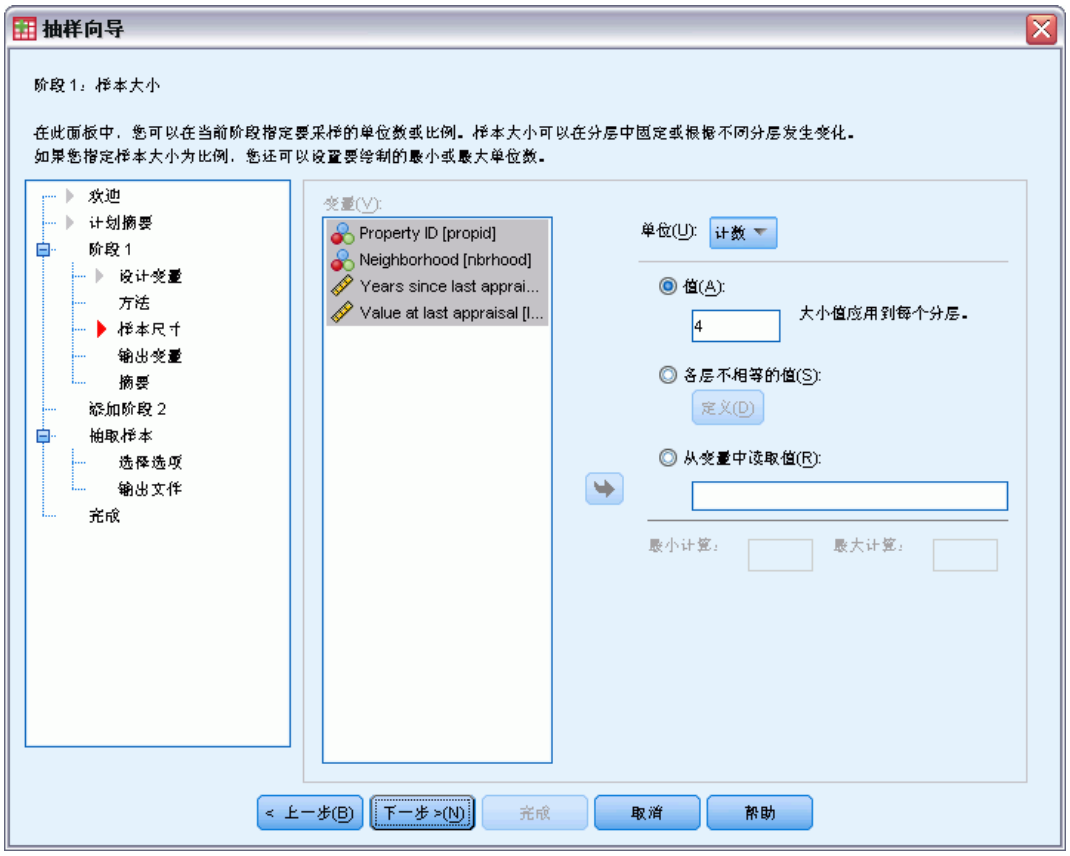
图片 13-2
抽样向导，“设计变量”步骤（阶段 1）



- 选择 County 作为分层变量。
- 选择 Township 作为聚类变量。
- 单击**下一步**，然后在“抽样方法”步骤中单击**下一步**。

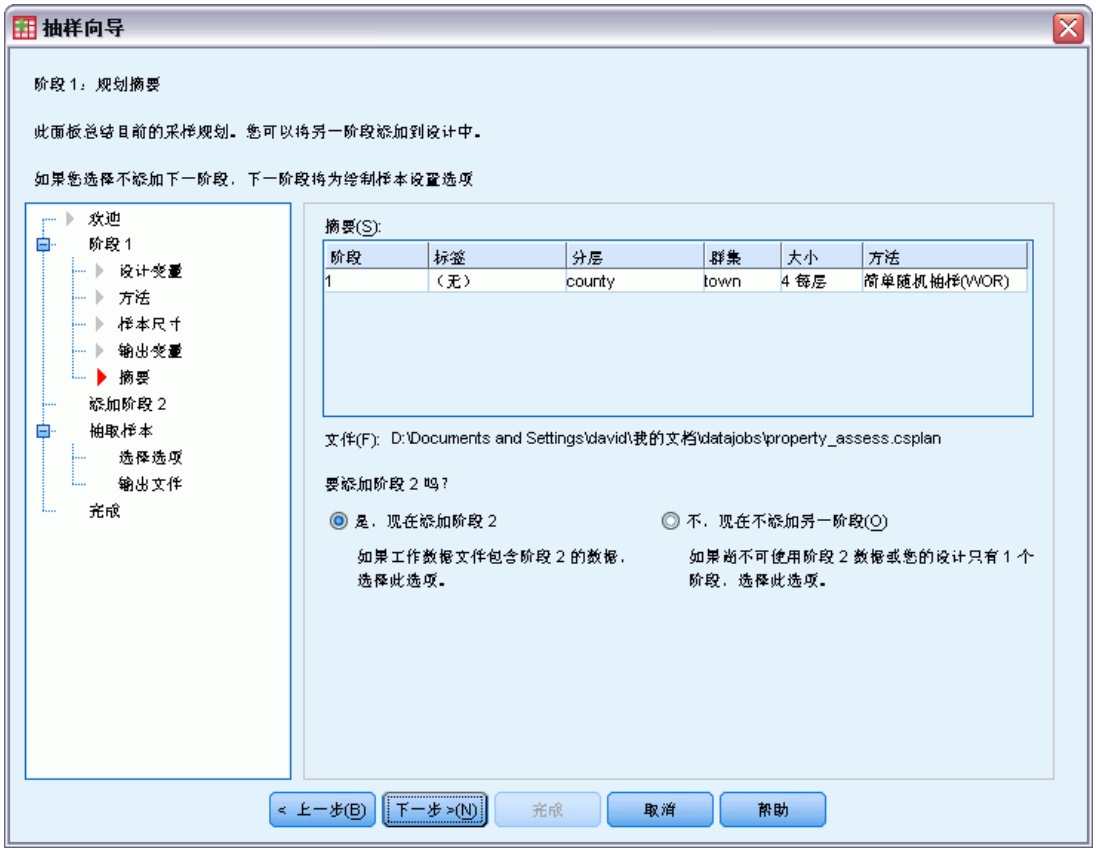
  此设计结构意味着为每个县抽取独立样本。这一阶段中，将使用缺省方法（简单随机抽样）将镇抽取为主抽样单元。

图片 13-3
抽样向导，“样本大小”步骤（阶段 1）



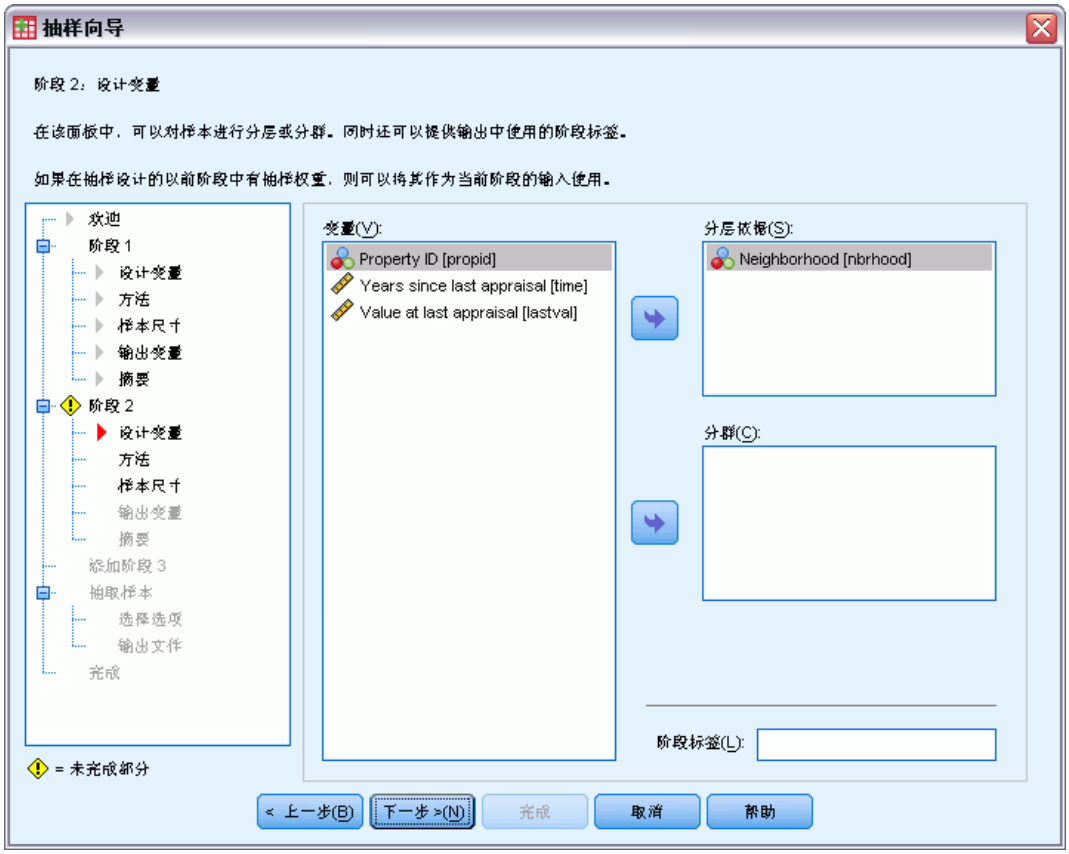
- 从?单元?下拉列表中选择**计数**。
- 键入 **4** 作为要在此阶段中选择的单元的数目。
- 单击**下一步**，然后在“输出变量”步骤中单击**下一步**。

图片 13-4
抽样向导，“计划摘要”步骤（阶段 1）



- 选择是，现在添加阶段 2。
- 单击下一步。

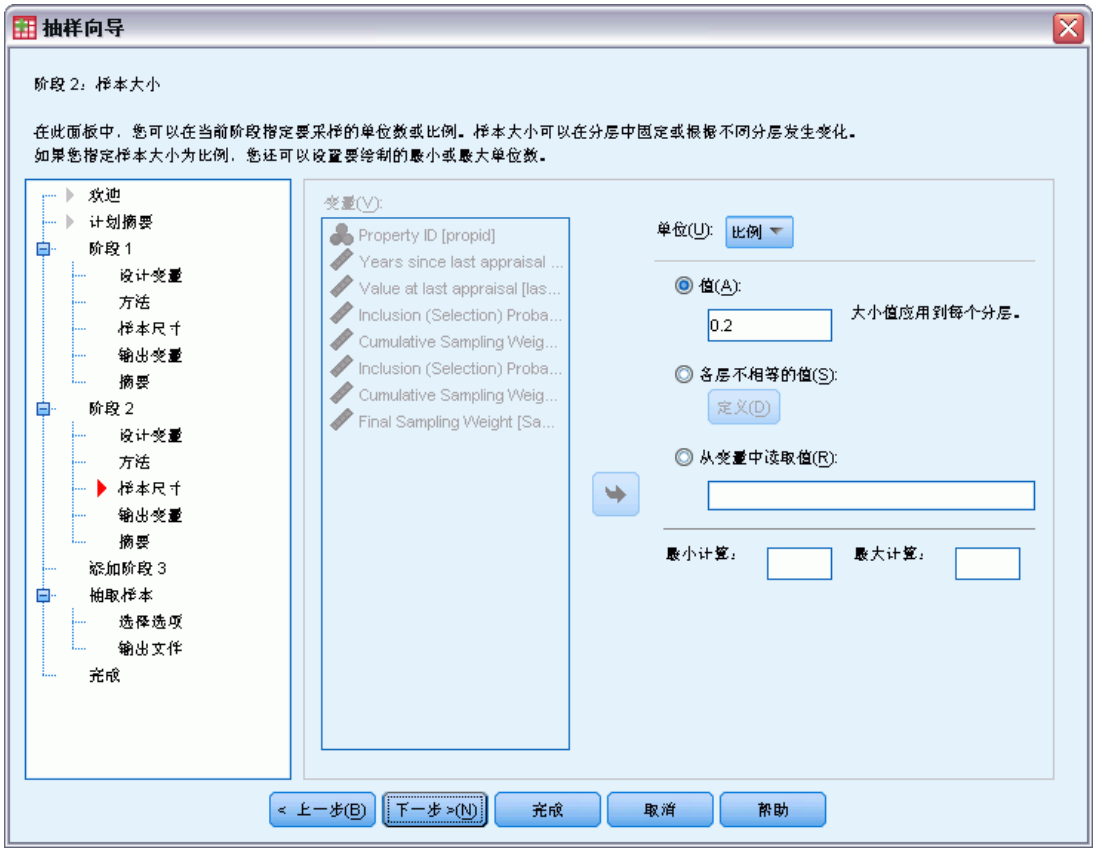
图片 13-5
抽样向导，“设计变量”步骤（阶段 2）



- 选择 Neighborhood 作为分层变量。
- 单击**下一步**，然后在“抽样方法”步骤中单击**下一步**。

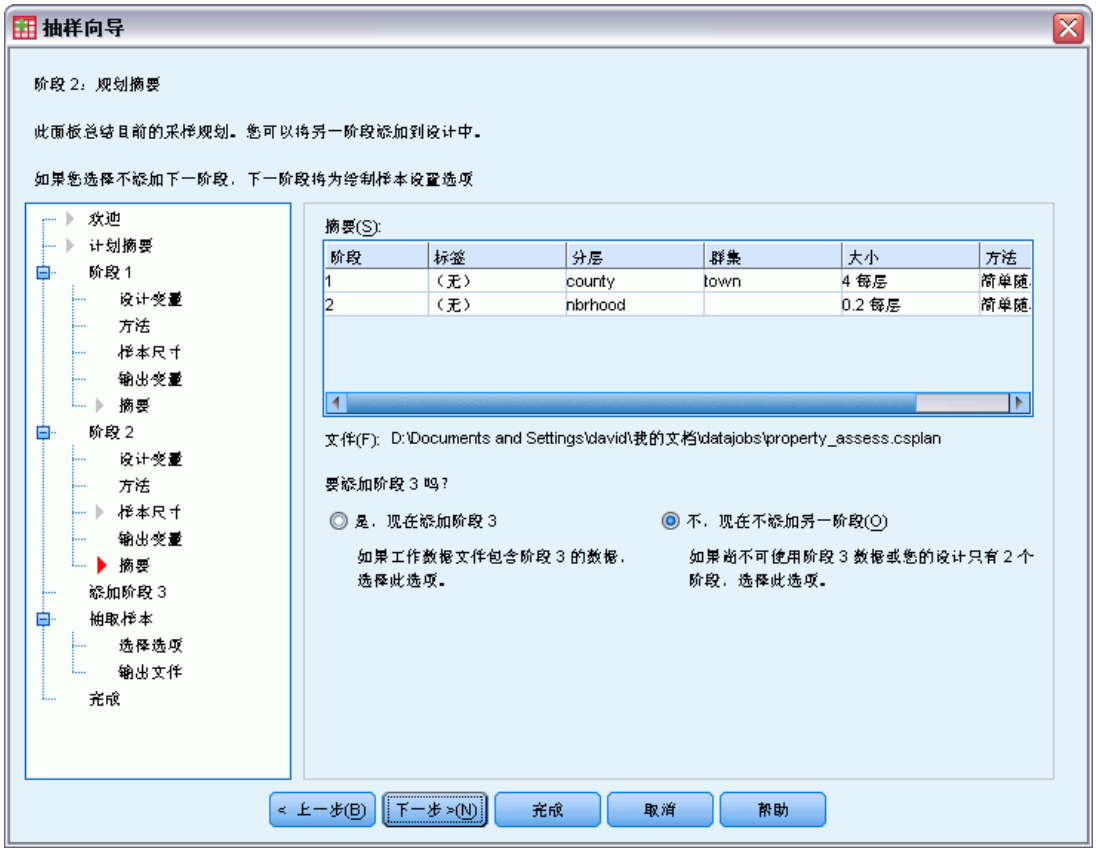
  此设计结构意味着会为阶段 1 抽取的镇的每个区抽取独立样本。这一阶段中，将使用简单随机抽样将资产抽取为主抽样单元。

图片 13-6
抽样向导，“样本大小”步骤（阶段 2）



- 从?单元?下拉列表中选择**比例**。
- 键入 **0.2** 作为要从每层抽样的单元的比例值。
- 单击**下一步**，然后在“输出变量”步骤中单击**下一步**。

图片 13-7
抽样向导，“计划摘要”步骤（阶段 2）



▶ 查看抽样设计，然后单击**下一步**。

图片 13-8
抽样向导，“抽取样本：选择选项”步骤

- 对于要使用的随机种子类型，选择**定制值**，并键入 **241972** 作为该值。

  使用定制值允许您精确重现此示例的结果。

- 单击**下一步**，然后在“抽取样本：输出文件”步骤中单击**下一步**。

图片 13-9
抽样向导，“完成”步骤

▶ 单击完成。

这些选择会生成抽样计划文件 property_assess.csplan 并根据该计划抽取样本。

计划摘要

图片 13-10
计划摘要

			阶段 1	阶段 2
设计变量	分层	1	County	Neighborhood
	群集	1	Township	
样本信息	选择方法		简单无替换随机抽样	简单无替换随机抽样
	已采样单位数量		4	
	创建或修改的变量	分阶段包含(选择)概率	Inclusion Probability_1_	Inclusion Probability_2_
		分阶段累积样本权重	Sample Weight Cumulative_1_	Sample Weight Cumulative_2_
	已采样单位百分比			.2
分析信息	估计量假设		无替换等概率抽样	无替换等概率抽样
	包含概率		从变量 Inclusion Probability_1_ 获得	从变量 Inclusion Probability_2_ 获得

规划文件: c:\property_assess.csplan
权重变量: SampleWeight_Final_

通过该摘要表可以复查您的抽样计划，这对于确保计划体现您的意图非常有用。

抽样摘要

图片 13-11
阶段摘要

	已采样单位数量		已采样单位百分比	
County	必需	实际	必需	实际
Eastern	4	4	44.4%	44.4%
Central	4	4	57.1%	57.1%
Western	4	4	25.0%	25.0%
Northern	4	4	44.4%	44.4%
Southern	4	4	50.0%	50.0%

规划文件: c:\property_assess.csplan

通过此摘要表可以复查抽样的第一阶段，这对于检查抽样是否按计划进行非常有用。按照要求，从每县抽样 4 个镇。

图片 13-12
阶段摘要

County	Township	Neighborhood	已采样单位数量 必需	已采样单位数量 实际	已采样单位百分比 必需	已采样单位百分比 实际
Eastern	2	8	4	4	20.0%	19.0%
		9	14	14	20.0%	20.6%
		10	7	7	20.0%	18.9%
		11	14	14	20.0%	20.0%
	6	36	13	13	20.0%	20.3%
		37	14	14	20.0%	20.6%
		38	13	13	20.0%	20.6%
	7	43	12	12	20.0%	20.7%
		44	11	11	20.0%	19.6%
		45	11	11	20.0%	20.8%
		46	13	13	20.0%	20.0%
	9	57	13	13	20.0%	20.6%
		58	5	5	20.0%	18.5%
		59	11	11	20.0%	19.3%
		60	13	13	20.0%	19.4%
Central	22	148	9	9	20.0%	19.6%
		149	8	8	20.0%	20.0%

通过此摘要表（此处显示的是表的顶部）可以复查抽样的第二阶段。该表对于检查抽样是否按计划进行也非常有用。按照要求，从第一阶段抽取的每个镇的每个区抽样大约 20% 的资产。

样本结果

图片 13-13
带有样本结果的数据编辑器

	propid	nbrhood	town	county	time	lastval	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	SampleWeight_Final_
273	13661.00	171	25	Central	7	152.20	.	.	.	.	.
274	13668.00	171	25	Central	6	104.30	.	.	.	.	.
275	13688.00	172	25	Central	6	203.00	.	.	.	.	.
276	13690.00	172	25	Central	7	201.40	.	.	.	.	.
277	13691.00	172	25	Central	6	188.30	.	.	.	.	.
278	13699.00	172	25	Central	6	207.00	.	.	.	.	.
279	13703.00	172	25	Central	6	179.90	0.57	1.75	0.19	9.19	9.19
280	13709.00	172	25	Central	6	85.30	0.57	1.75	0.19	9.19	9.19
281	13712.00	172	25	Central	6	162.30	.	.	.	.	.
282	13719.00	172	25	Central	7	154.50	.	.	.	.	.
283	14563.00	183	27	Central	5	158.10	0.57	1.75	0.20	8.62	8.62
284	14566.00	183	27	Central	6	104.60					

数据视图 变量视图

可在数据编辑器中看到抽样结果。 五个新变量保存到了工作文件中，分别表示每个阶段的包含概率和累积抽样权重，以及最终抽样权重。

- 将具有这些变量值的个案选入样本。
- 不选择具有这些变量的系统缺失值的个案。

该机构现在将使用其资源来收集样本中所选资产的当前评估值。一旦这些评估值可用，即可使用抽样计划 property_assess.csplan 提供抽样指定项，通过复杂样本分析过程对样本进行处理。

从部分抽样框架获取样本

某公司对构建和销售高质量调查信息数据库感兴趣。调查样本应具有代表性并可高效执行，因此使用复杂抽样方法。完整抽样设计需要以下结构：

阶段	分层	群集
1	地区	省
2	区	市
3	子区	

在第三阶段中，家庭是主抽样单元，并将对所选家庭进行调查。但是，由于只有城市级别的信息才便于获得，该公司计划现在执行设计的前两个阶段，然后从抽样城市收集子区和家庭数量的信息。有关城市级别的可用信息收集在 demo_cs_1.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件. 请注意，该文件包含变量子区，该变量的所有值都为 1。这是搵rue使用复杂样本抽样向导指定完整的复杂抽样设计，然后进行前两个阶段的抽取。

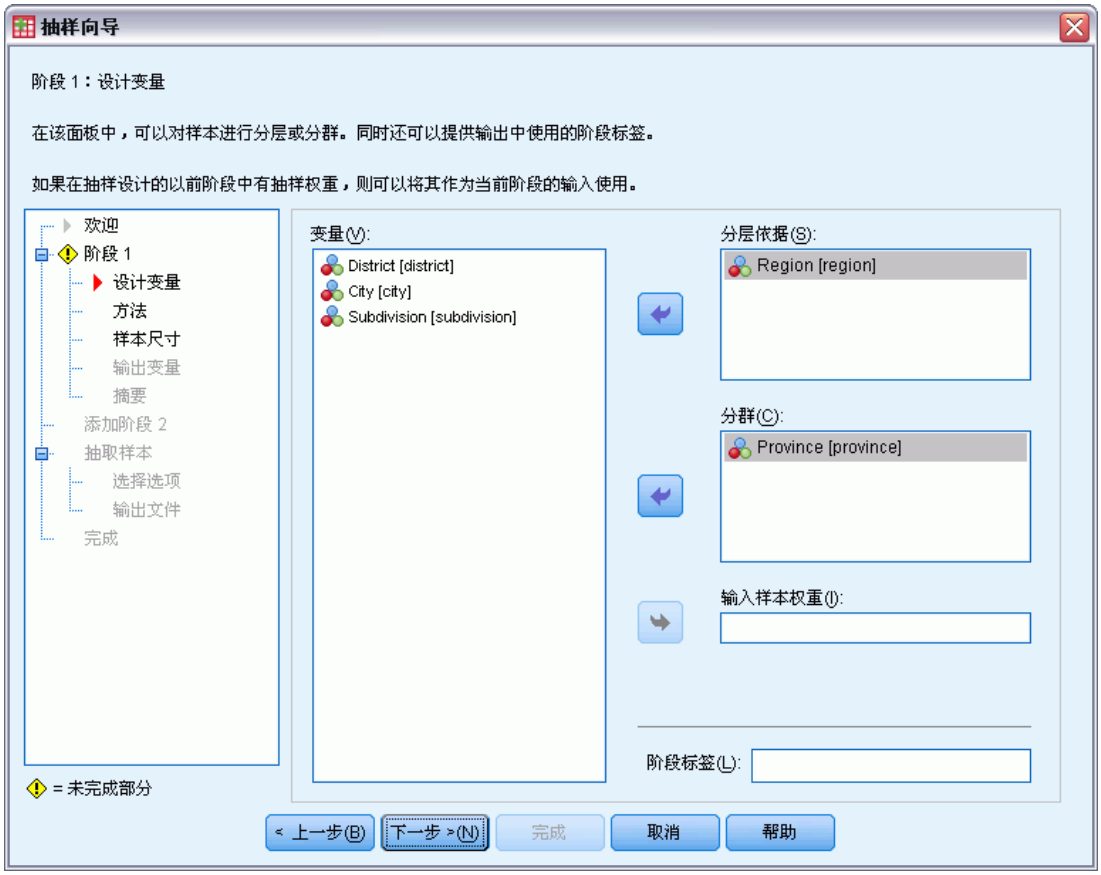
使用该向导从第一部分框架抽样

▶ 要运行复杂样本抽样向导，请从菜单中选择：
分析 > 复杂抽样 > 选择样本...

图片 13-14
抽样向导，“欢迎”步骤

- 选择**设计样本**，浏览到要保存文件的位置，并输入 demo.csplan 作为计划文件的名称。
- 单击**下一步**。

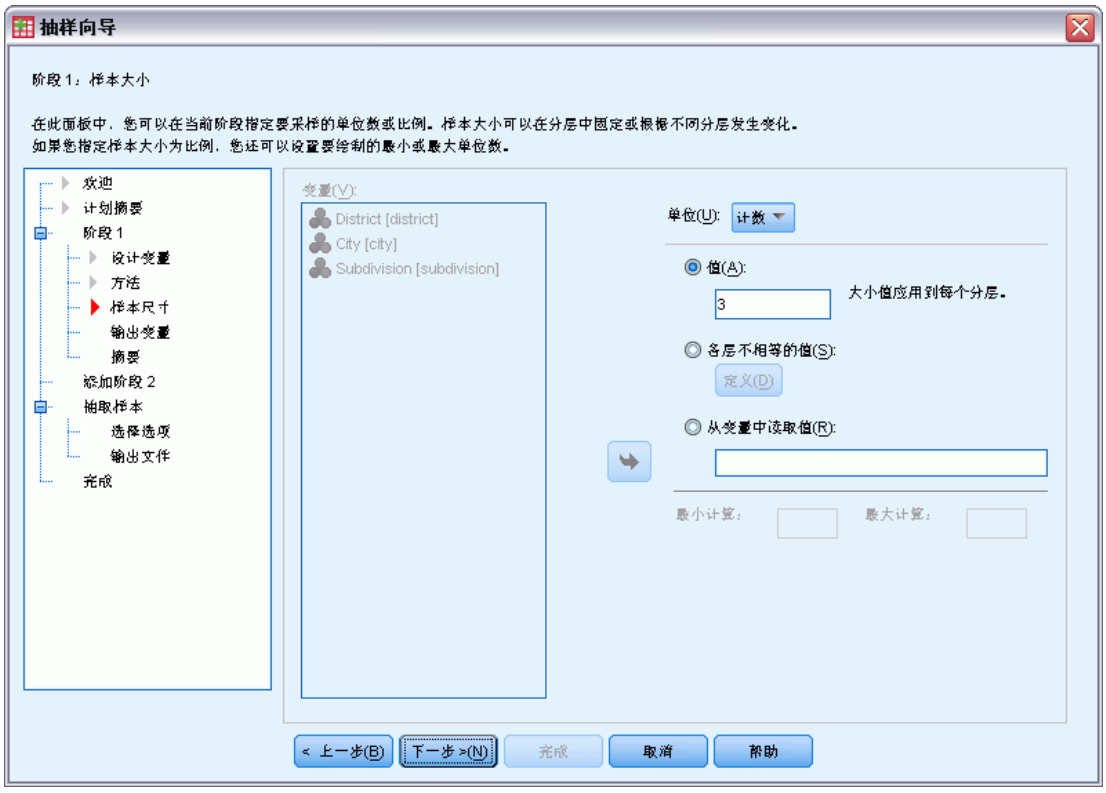
图片 13-15
抽样向导，“设计变量”步骤（阶段 1）



- 选择 Region 作为分层变量。
- 选择 Province 作为聚类变量。
- 单击**下一步**，然后在“抽样方法”步骤中单击**下一步**。

  此设计结构意味着为每个区域抽取独立样本。这一阶段中，使用缺省方法（简单随机抽样）将省抽取为主抽样单元。

图片 13-16
抽样向导，“样本大小”步骤（阶段 1）

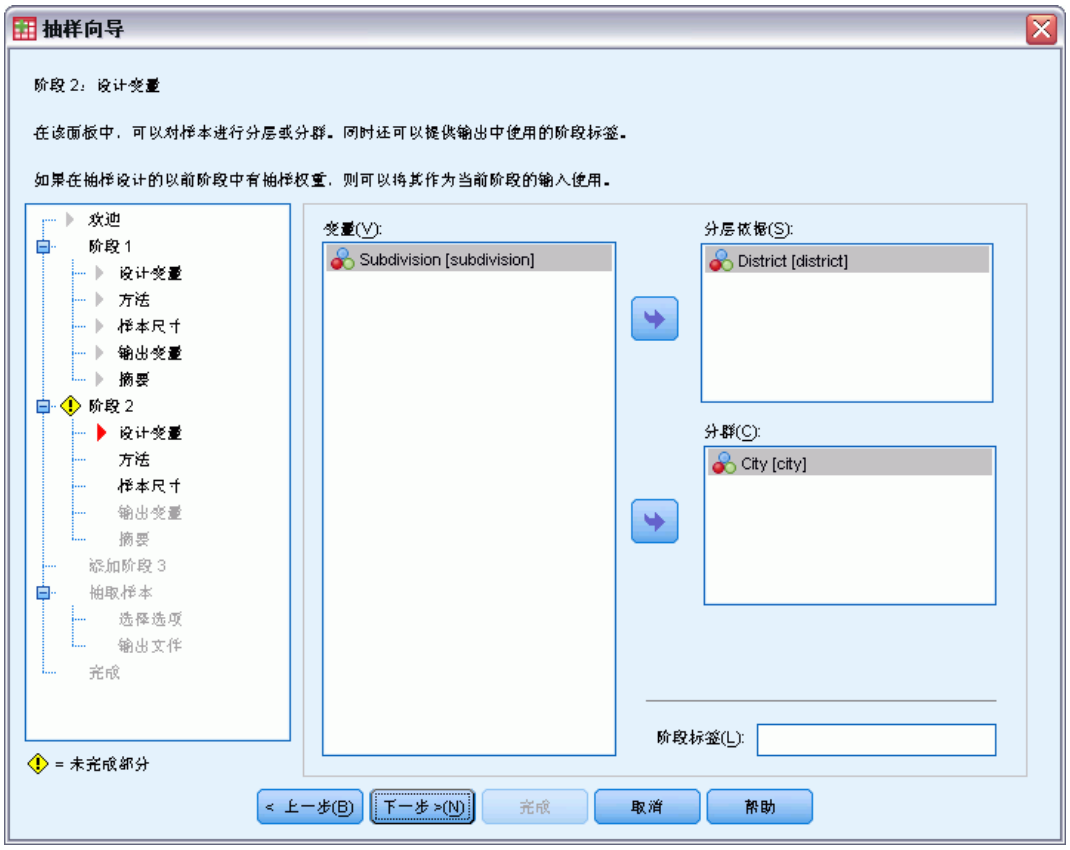


- 从?单元?下拉列表中选择**计数**。
- 键入 **3** 作为要在此阶段中选择的单元的数目。
- 单击**下一步**，然后在“输出变量”步骤中单击**下一步**。

图片 13-17
抽样向导，“计划摘要”步骤（阶段 1）

- 选择是，现在添加阶段 2。
- 单击下一步。

图片 13-18
抽样向导，“设计变量”步骤（阶段 2）



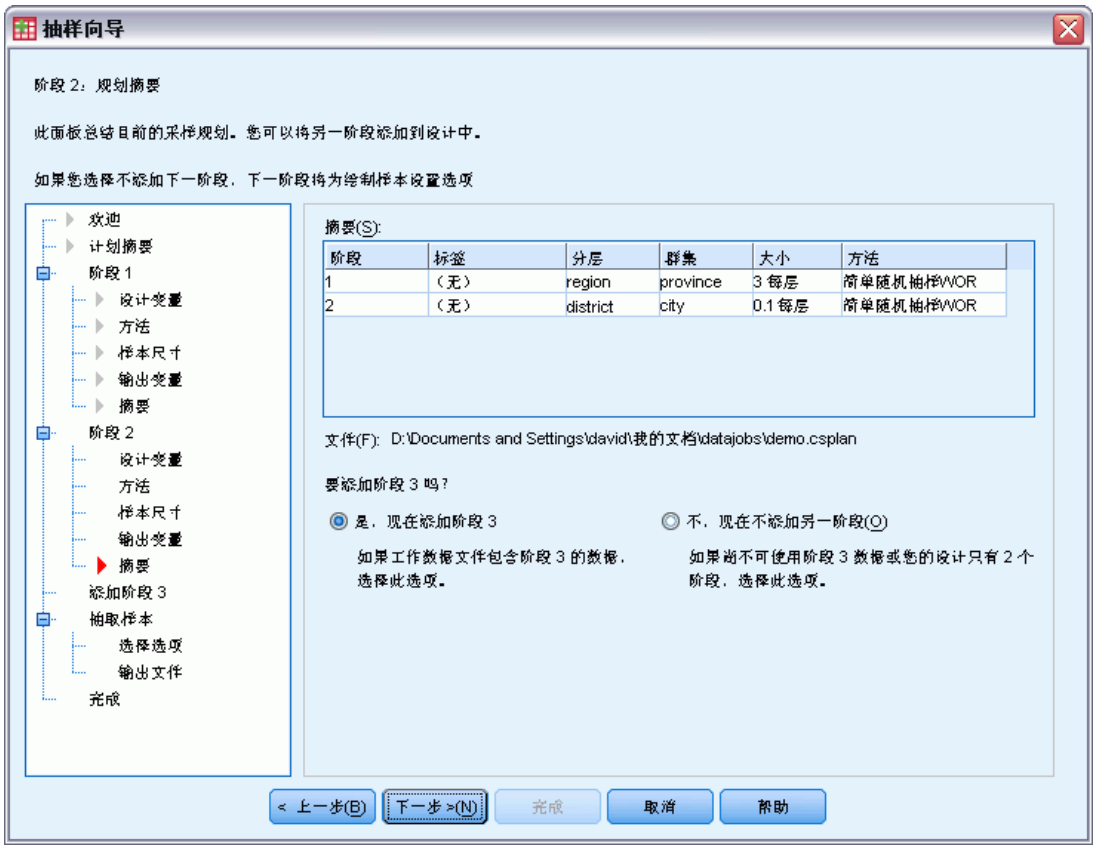
- 选择 District 作为分层变量。
- 选择 City 作为聚类变量。
- 单击**下一步**，然后在“抽样方法”步骤中单击**下一步**。

  此设计结构意味着为每个地区抽取独立样本。这一阶段中，使用缺省方法（简单随机抽样）将城市抽取为主抽样单元。

图片 13-19
抽样向导，“样本大小”步骤（阶段 2）

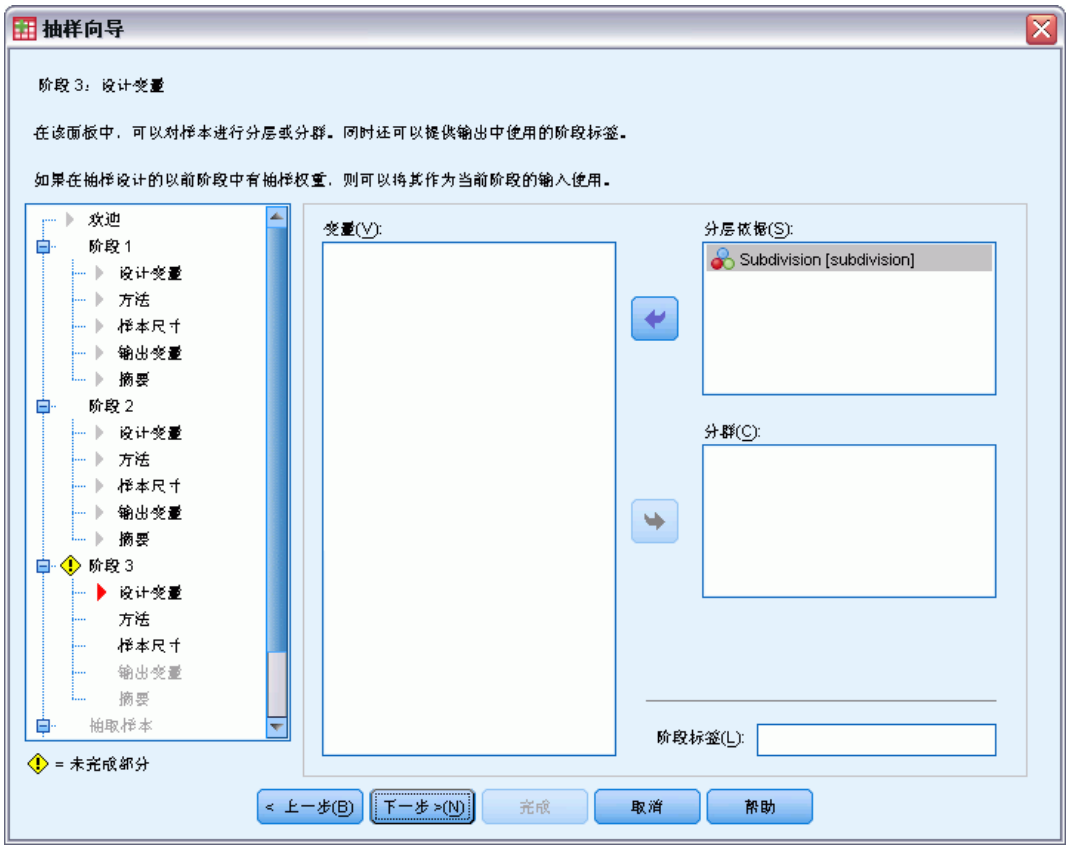
- 从?单元?下拉列表中选择比例。
- 键入 0.1 作为要从每层抽样的单元的比例值。
- 单击下一步，然后在“输出变量”步骤中单击下一步。

图片 13-20
抽样向导，“计划摘要”步骤（阶段 2）



- 选择是，现在添加阶段 3。
- 单击下一步。

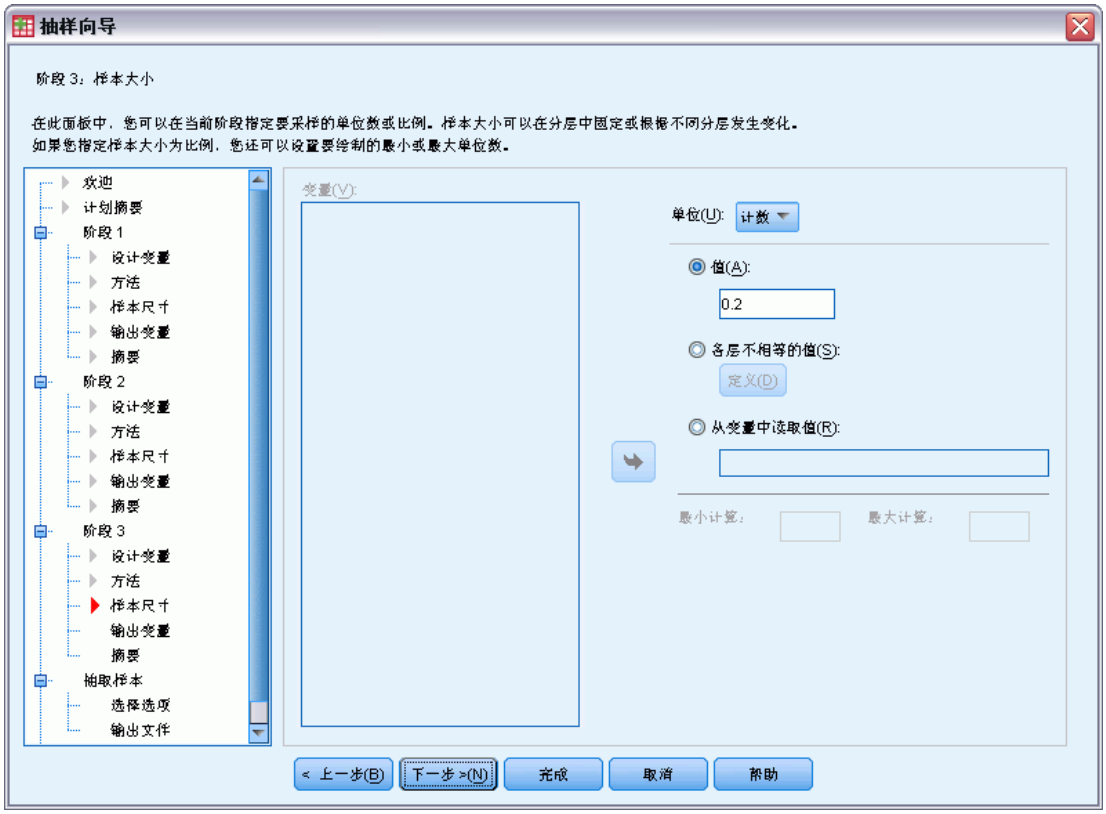
图片 13-21
抽样向导，“设计变量”步骤（阶段 3）



- 选择子区作为分层变量。
- 单击**下一步**，然后在“抽样方法”步骤中单击**下一步**。

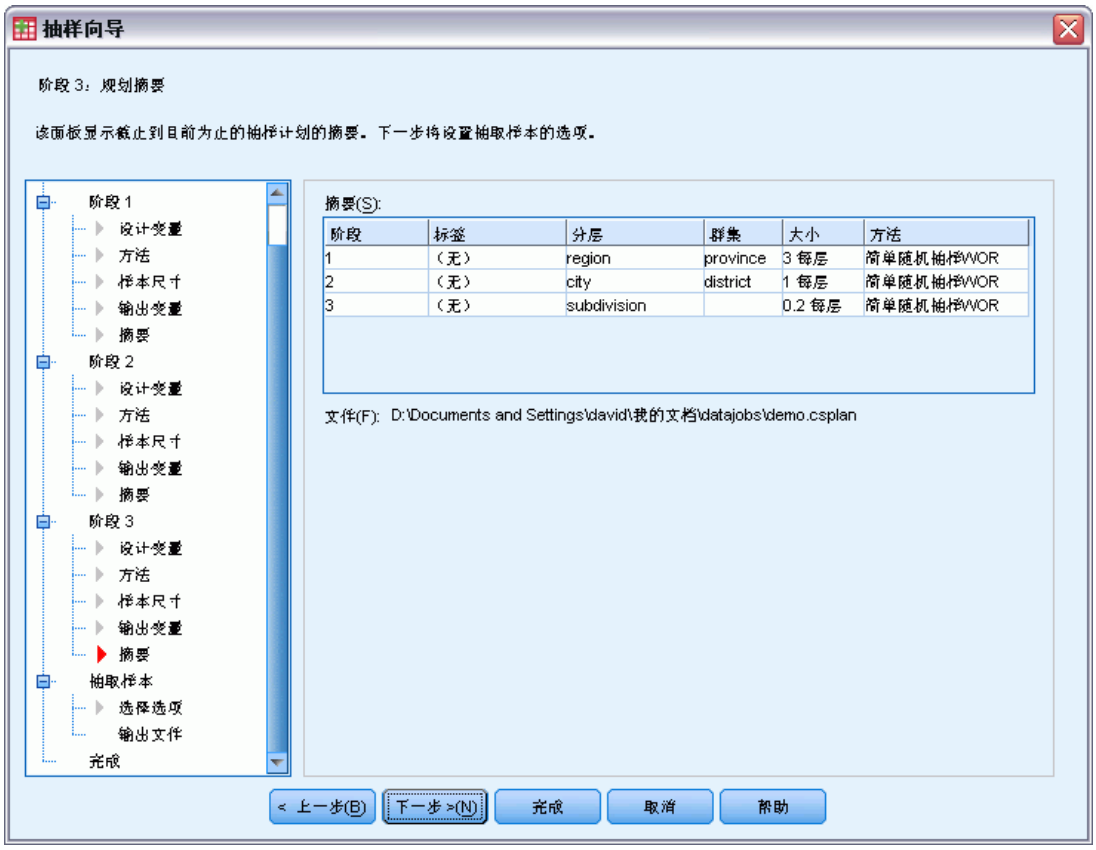
  此设计结构意味着为每个子区抽取独立样本。这一阶段中，使用缺省方法（简单随机抽样）将家庭单元抽取为主抽样单元。

图片 13-22
抽样向导，“样本大小”步骤（阶段 3）



- 从?单元?下拉列表中选择**比例**。
- 键入 **0.2** 作为要在此阶段中选择的单元的比例值。
- 单击**下一步**，然后在“输出变量”步骤中单击**下一步**。

图片 13-23
抽样向导，“计划摘要”步骤（阶段 3）



▶ 查看抽样设计，然后单击**下一步**。

图片 13-24
抽样向导，“抽取样本选择选项”步骤

- 选择 1, 2 作为现在要抽样的阶段。
- 对于要使用的随机种子类型，选择**定制值**，并键入 241972 作为该值。

  使用定制值允许您精确重现此示例的结果。
- 单击**下一步**，然后在“抽取样本：输出文件”步骤中单击**下一步**。

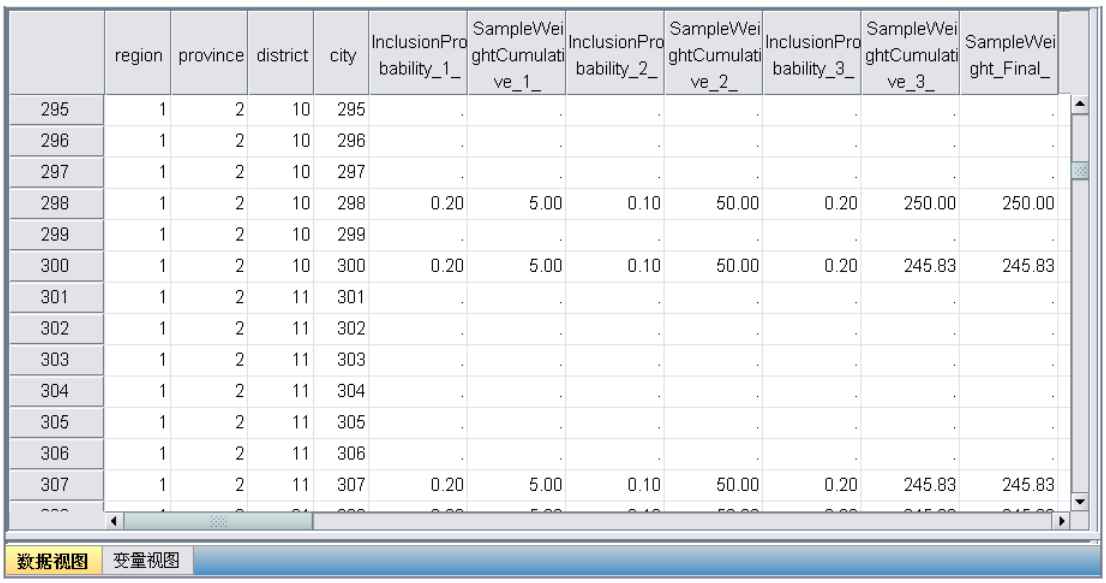
图片 13-25
抽样向导，“完成”步骤

▶ 单击完成。

这些选择会生成抽样计划文件 demo.csplan 并根据该计划的前两个阶段抽取样本。

样本结果

图片 13-26
带有样本结果的数据编辑器



	region	province	district	city	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	InclusionProbability_3_	SampleWeightCumulative_3_	SampleWeight_Final_
295	1	2	10	295	.	.	.	.	.	.	.
296	1	2	10	296	.	.	.	.	.	.	.
297	1	2	10	297	.	.	.	.	.	.	.
298	1	2	10	298	0.20	5.00	0.10	50.00	0.20	250.00	250.00
299	1	2	10	299	.	.	.	.	.	.	.
300	1	2	10	300	0.20	5.00	0.10	50.00	0.20	245.83	245.83
301	1	2	11	301	.	.	.	.	.	.	.
302	1	2	11	302	.	.	.	.	.	.	.
303	1	2	11	303	.	.	.	.	.	.	.
304	1	2	11	304	.	.	.	.	.	.	.
305	1	2	11	305	.	.	.	.	.	.	.
306	1	2	11	306	.	.	.	.	.	.	.
307	1	2	11	307	0.20	5.00	0.10	50.00	0.20	245.83	245.83

数据视图 变量视图

可在数据编辑器中看到抽样结果。 五个新变量保存到了工作文件中，分别表示每个阶段的包含概率和累积抽样权重，以及前两个阶段的“最终”抽样权重。

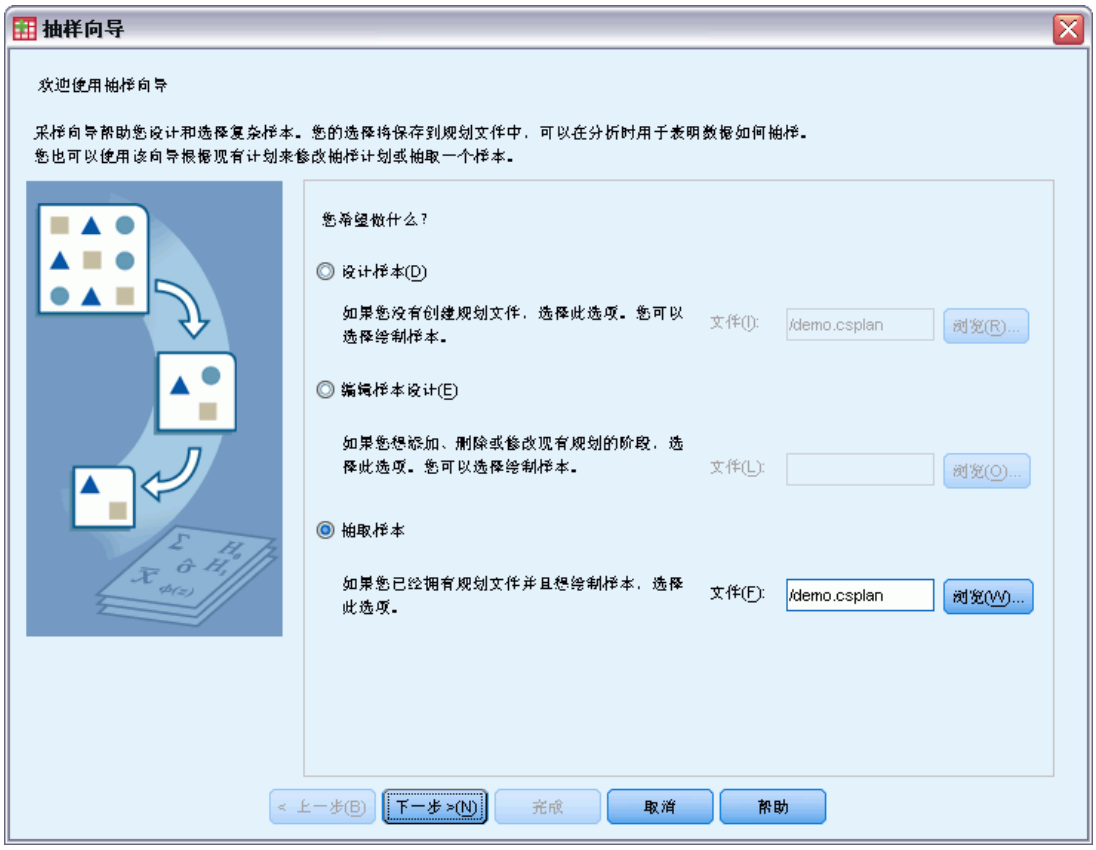
- 将具有这些变量值的城市选入样本。
- 不选择具有这些变量的系统缺失值的城市。

对于选择的每个城市，该公司调查其子区和家庭单元的信息并放置在 demo_cs_2.sav 中。使用此文件和抽样向导进行此设计第三阶段的抽样。

使用该向导从第二部分框架抽样

▶ 要运行复杂样本抽样向导，请从菜单中选择：
分析 > 复杂抽样 > 选择样本...

图片 13-27
抽样向导，“欢迎”步骤



- 选择**抽取样本**，浏览到保存计划文件的位置，并选择您创建的 demo.csplan 计划文件。
- 单击**下一步**。

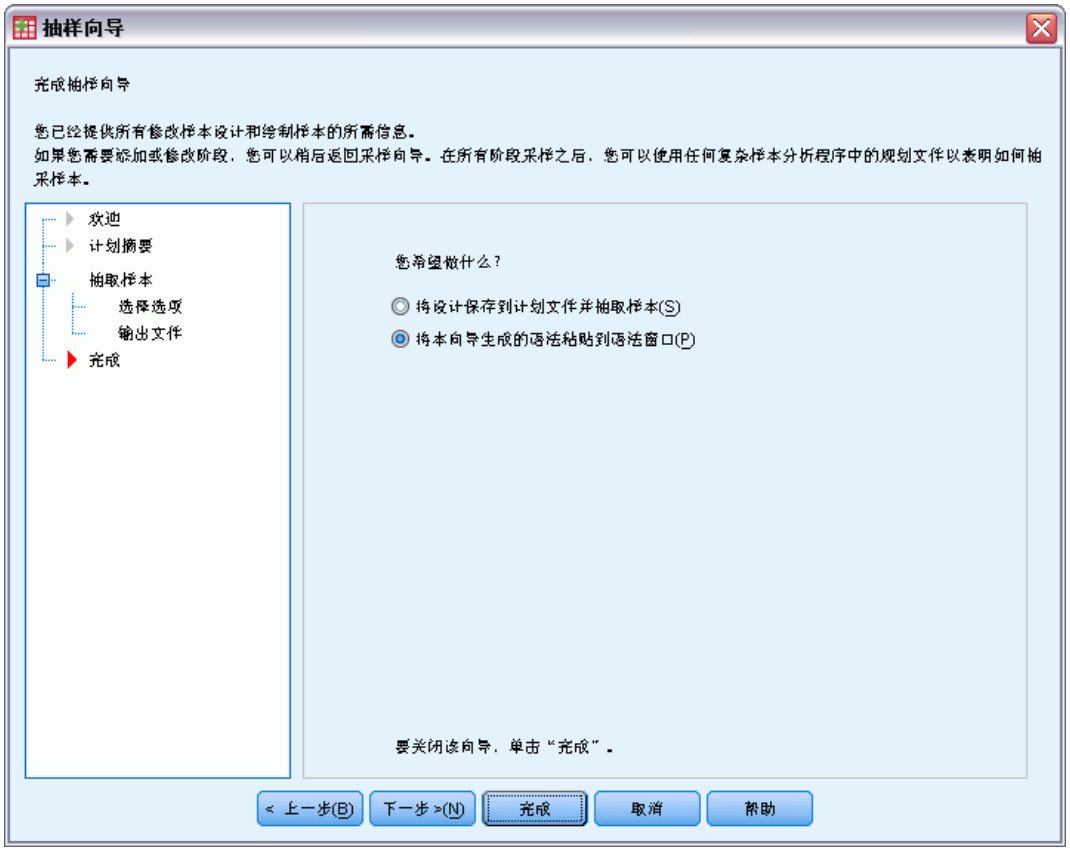
图片 13-28
抽样向导，“计划摘要”步骤（阶段 3）

- 选择 1，2 作为已抽样的阶段。
- 单击下一步。

图片 13-29
抽样向导，“抽取样本选择选项”步骤

- 对于要使用的随机种子类型，选择**定制值**，并键入 **4231946** 作为该值。
- 单击**下一步**，然后在“抽取样本：输出文件”步骤中单击**下一步**。

图片 13-30
抽样向导，“完成”步骤



- 选择将本向导生成的语法粘贴到语法窗口。
- 单击完成。

生成以下语法：

```
* Sampling Wizard.
CSSELECT
 /PLAN FILE='demo.csplan'
 /CRITERIA STAGES = 3 SEED = 4231946
 /CLASSMISSING EXCLUDE
 /DATA RENAMEVARS
 /PRINT SELECTION.
```

这种情况下打印抽样摘要会生成一个繁琐的表，会在输出浏览器中导致出现问题。要关闭抽样摘要的显示，请在 PRINT 子命令中以 CPS 替换 SELECTION。然后，在语法窗口中运行该语法。

这些选择将根据 demo.csplan 抽样计划的第三阶段抽取样本。

样本结果

图片 13-31
带有样本结果的数据编辑器

	city	subdivision	unit	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	InclusionProbability_3_	SampleWeightCumulative_3_	SampleWeight_Final_
14	190	946	94514	0.20	5.00	0.10	50.00	.	.	.
15	190	946	94515	0.20	5.00	0.10	50.00	.	.	.
16	190	946	94516	0.20	5.00	0.10	50.00	0.20	244.44	244.44
17	190	946	94517	0.20	5.00	0.10	50.00	.	.	.
18	190	946	94518	0.20	5.00	0.10	50.00	.	.	.
19	190	946	94519	0.20	5.00	0.10	50.00	.	.	.
20	190	946	94520	0.20	5.00	0.10	50.00	.	.	.
21	190	946	94521	0.20	5.00	0.10	50.00	.	.	.
22	190	946	94522	0.20	5.00	0.10	50.00	.	.	.
23	190	946	94523	0.20	5.00	0.10	50.00	.	.	.
24	190	946	94524	0.20	5.00	0.10	50.00	0.20	244.44	244.44
25	190	946	94525	0.20	5.00	0.10	50.00	.	.	.
26	190	946	94526	0.20	5.00	0.10	50.00	.	.	.

数据视图 变量视图

可在数据编辑器中看到抽样结果。三个新变量保存到了工作文件中，分别表示第三阶段的包含概率和累积抽样权重，以及最终抽样权重。这些新权重考虑了前两个阶段的抽样过程中计算得到的权重。

- 将具有这些变量值的单元选入样本。
- 不选择具有这些变量的系统缺失值的单元。

现在，该公司将使用其资源来获取样本中所选的家庭单元的调查信息。收集了调查信息之后，即可使用抽样计划 demo.csplan 提供抽样指定项，通过复杂样本分析过程对样本进行处理。

以与大小成正比的概率抽样 (PPS)

议员在向立法院提交某项法案之前想了解公众是否支持该法案，以及对该法案的支持与选民人群统计信息有何关联。民意测验专家根据复杂抽样设计并实施了一些采访。

注册选民的列表收集在 poll_cs.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件. 使用复杂样本抽样向导选择样本进行下一步分析。

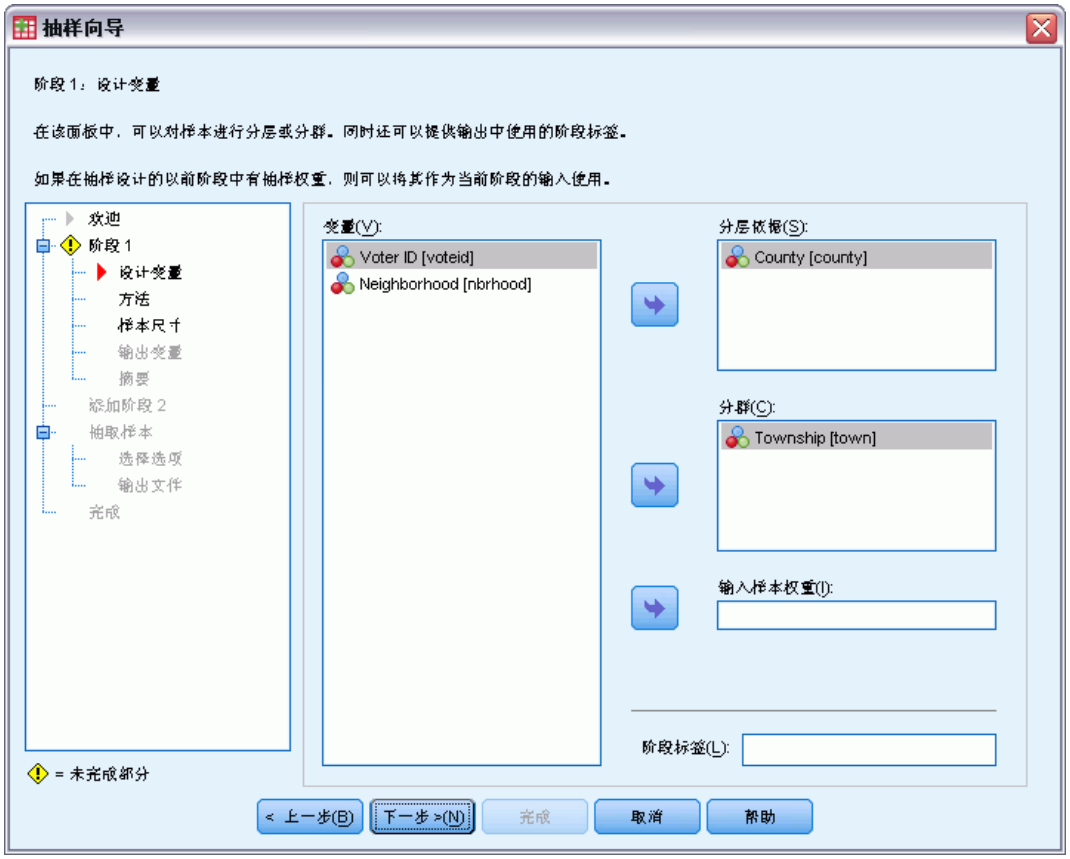
使用向导

▶ 要运行复杂样本抽样向导，请从菜单中选择：
分析 > 复杂抽样 > 选择样本...

图片 13-32
抽样向导，“欢迎”步骤

- 选择**设计样本**，浏览到要保存文件的位置，并输入 poll.csplan 作为计划文件的名称。
- 单击**下一步**。

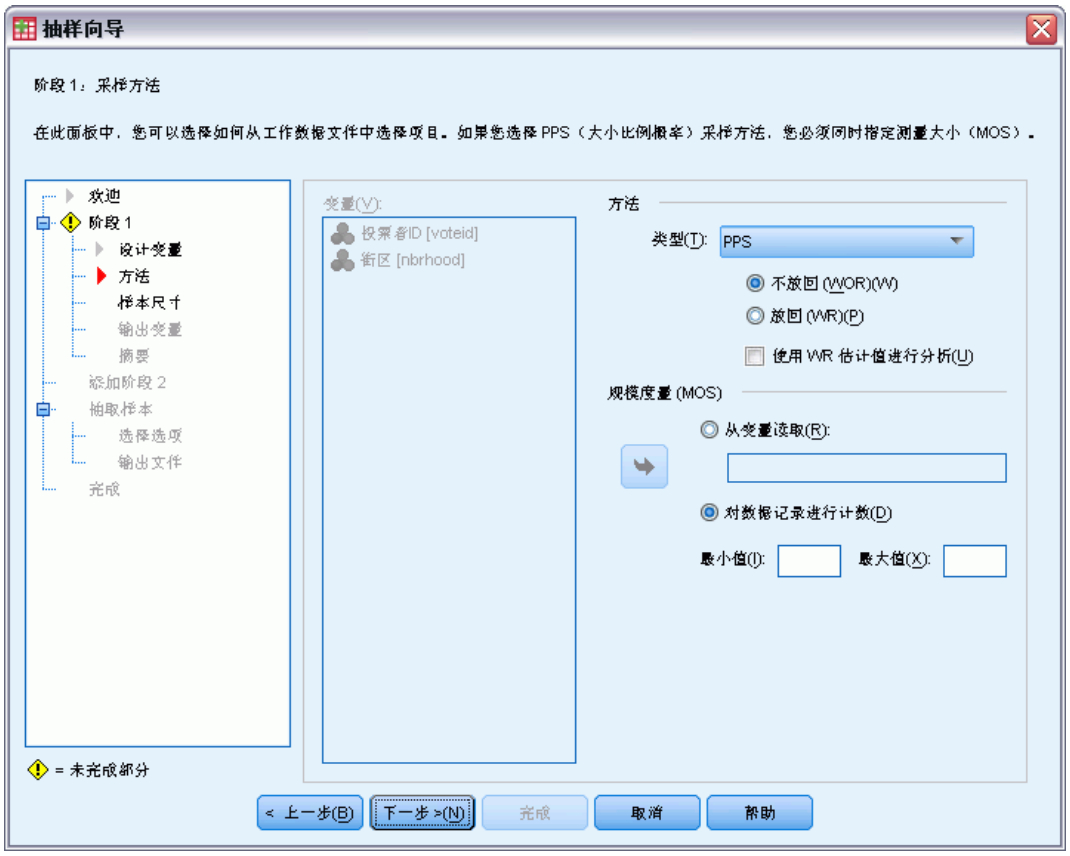
图片 13-33
抽样向导，“设计变量”步骤（阶段 1）



- 选择 County 作为分层变量。
- 选择 Township 作为聚类变量。
- 单击**下一步**。

  此设计结构意味着为每个县抽取独立样本。在这一阶段中，镇抽取为主抽样单元。

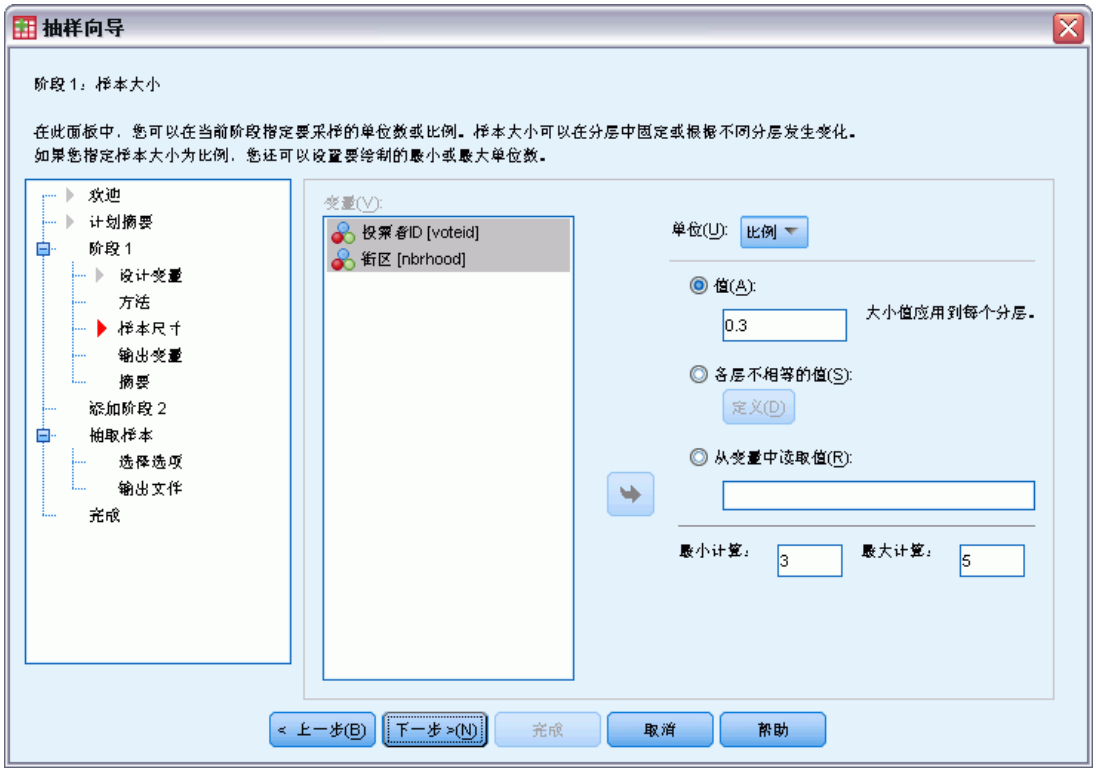
图片 13-34
抽样向导，“抽样方法”步骤（阶段 1）



- 选择 PPS 作为抽样方法。
- 选择 Count data records 作为规模度量。
- 单击下一步。

  在每个县中，都以和每镇记录数成正比的概率采用不放回方式抽取镇。使用 PPS 方法生成镇的联合抽样概率；您将在“输出文件”步骤中指定这些值的保存位置。

图片 13-35
抽样向导，“样本大小”步骤（阶段 1）



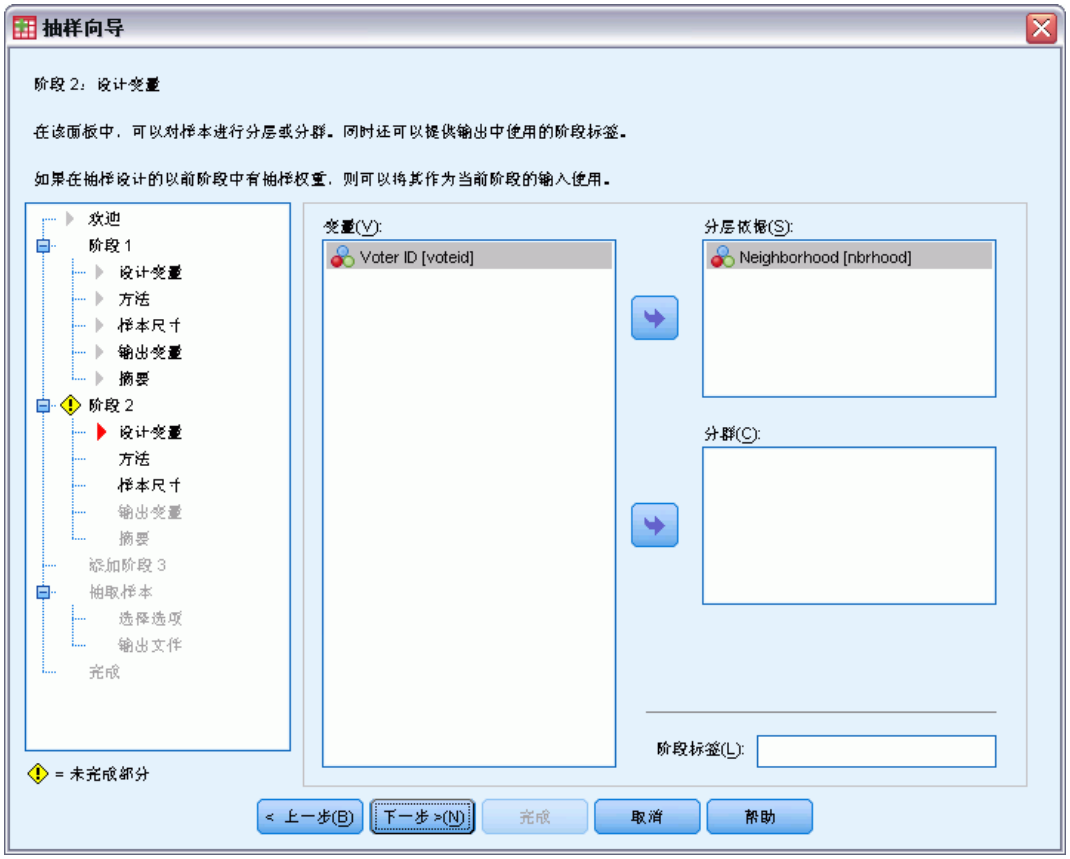
- 从?单元?下拉列表中选择**比例**。
- 键入 **0.3** 作为要在此阶段中从每县选择的镇的比例值。

  西县的立法者指出，他们县的镇比其他县少。为了确保足够的代表性，他们希望从每县最少抽样 3 个镇。
- 键入 **3** 作为要选择的最小镇数，键入 **5** 作为最大值。
- 单击**下一步**，然后在“输出变量”步骤中单击**下一步**。

图片 13-36
抽样向导，“计划摘要”步骤（阶段 1）

- 选择是，现在添加阶段 2。
- 单击下一步。

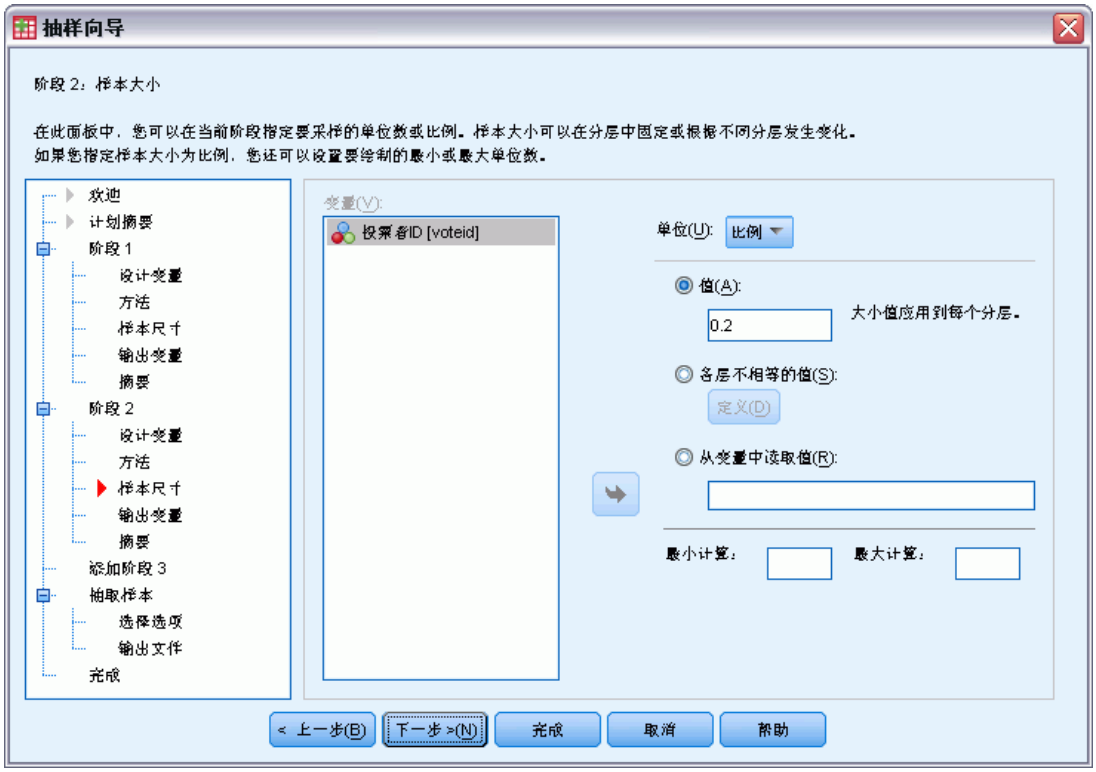
图片 13-37
抽样向导，“设计变量”步骤（阶段 2）



- 选择 Neighborhood 作为分层变量。
- 单击**下一步**，然后在“抽样方法”步骤中单击**下一步**。

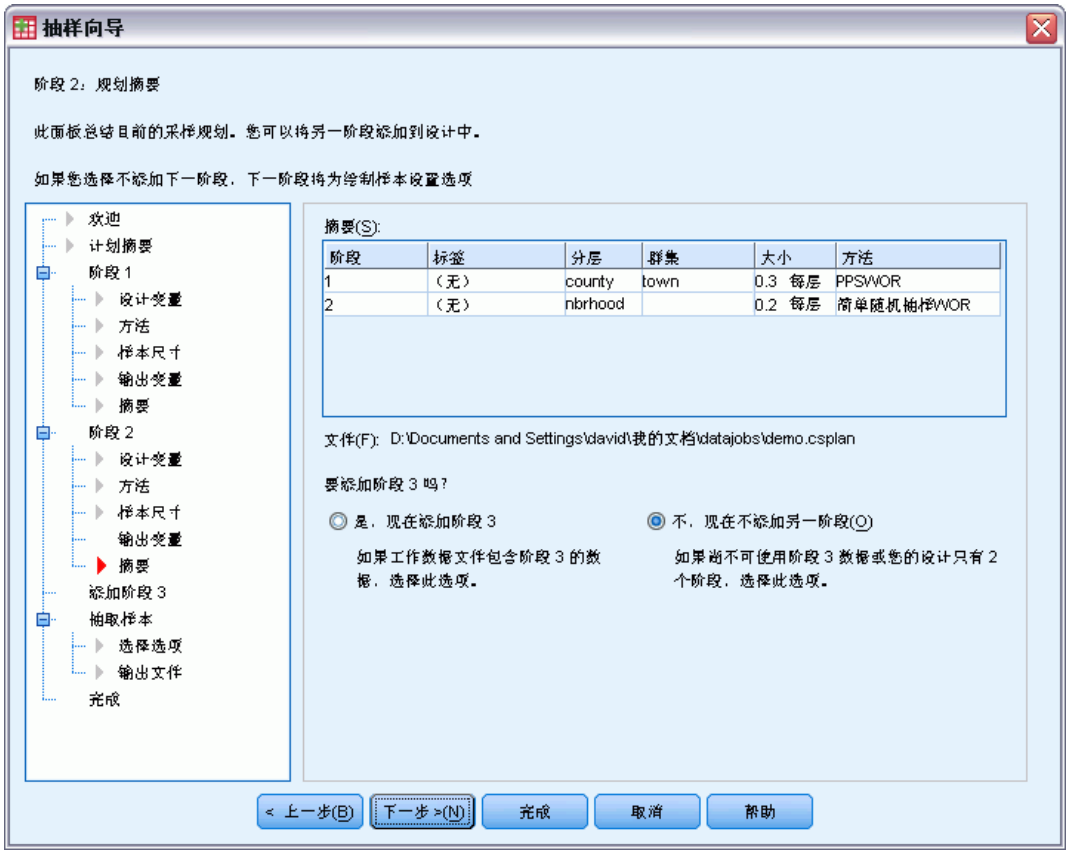
  此设计结构意味着会为阶段 1 抽取的镇的每个镇区抽取独立样本。这一阶段中，使用简单随机不放回方式抽样将选民抽取为主抽样单元。

图片 13-38
抽样向导，“样本大小”步骤（阶段 2）



- 从?单元?下拉列表中选择**比例**。
- 键入 .2 作为要从每层抽样的单元的比例值。
- 单击**下一步**，然后在“输出变量”步骤中单击**下一步**。

图片 13-39
抽样向导，“计划摘要”步骤（阶段 2）



▶ 查看抽样设计，然后单击**下一步**。

图片 13-40
抽样向导，“抽取样本选择选项”步骤

- 对于要使用的随机种子类型，选择**定制值**，并键入 592004 作为该值。

  使用定制值允许您精确重现此示例的结果。

- 单击**下一步**。

图片 13-41
抽样向导，“抽取样本选择选项”步骤

- 选择将样本保存到新的数据集，并键入 poll_cs_sample 作为数据集的名称。
- 浏览到要保存联合概率的位置，并输入 poll_jointprob.sav 作为联合概率文件的名称。
- 单击下一步。

图片 13-42
抽样向导，“完成”步骤

▶ 单击完成。

这些选择将生成抽样计划文件 poll.csplan 并根据该计划抽取样本，将样本结果保存到新数据集 poll_cs_sample 中，将联合概率文件保存到外部数据文件 poll_jointprob.sav 中。

计划摘要

图片 13-43
计划摘要

			阶段 1	阶段 2
设计变量	分层	1	县	街区
	群集	1	城镇	
样本信息	选择方法		无替换 PPS 抽样	简单无替换随机抽样
	度量大小		从数据获得	
	已采样单位百分比		.3	.2
	最小已采样单位数量		3	
	最大已采样单位数量		5	
	创建或修改的变量	分阶段包含（选择）概率	Inclusion Probability_1_	Inclusion Probability_2_
		分阶段累积样本权重	Sample Weight Cumulative_1_	Sample Weight Cumulative_2_
分析信息	估计量假设		无替换不等概率抽样（使用联合包含概率）	无替换等概率抽样
	包含概率		从变量 Inclusion Probability_1_ 获得	从变量 Inclusion Probability_2_ 获得

规划文件: c:\poll.csplan
权重变量: SampleWeight_Final_

通过该摘要表可以复查您的抽样计划，这对于确保计划体现您的意图非常有用。

抽样摘要

图片 13-44
阶段摘要

	已采样单位数量		已采样单位百分比	
县	必需	实际	必需	实际
东部	4	4	30.0%	30.8%
中部	4	4	30.0%	30.8%
西部	3	3	30.0%	50.0%
北部	5	5	30.0%	33.3%
南部	3	3	30.0%	50.0%

规划文件: c:\poll.csplan

通过此摘要表可以复查抽样的第一阶段，这对于检查抽样是否按计划进行非常有用。您请求的是抽取县中 30% 的镇；除了西县和南县之外，实际抽样比例接近 30%。这是因为这些县都只有六个镇，而且还指定了每县应抽取的最少镇数为 3。

图片 13-45
阶段摘要

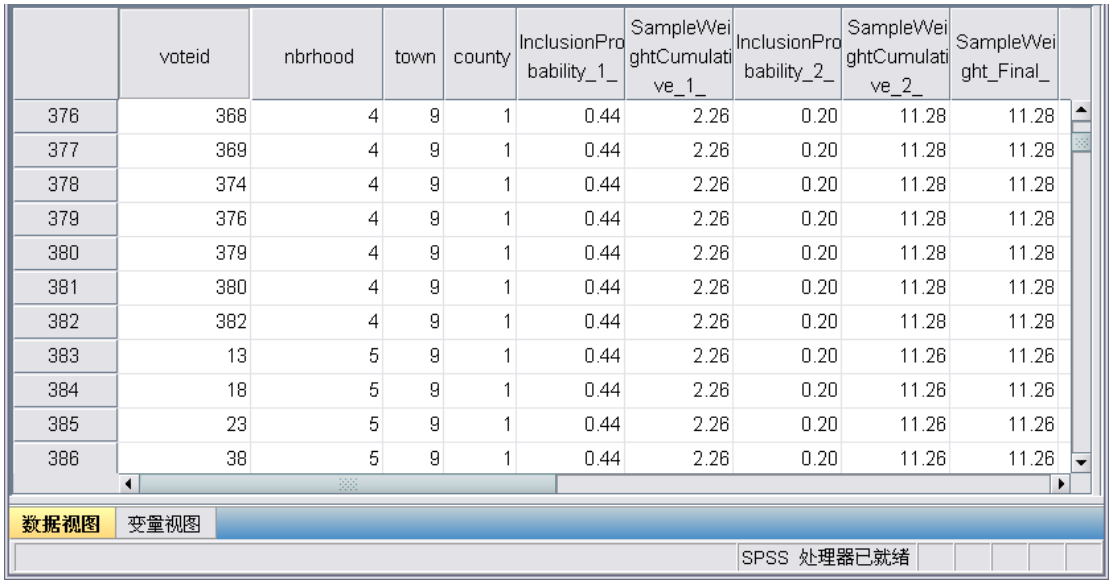
			已采样单位数量		已采样单位百分比	
县	城镇	街区	必需	实际	必需	实际
东部	9	1	49	49	20.0%	19.9%
		2	143	143	20.0%	20.0%
		3	113	113	20.0%	20.0%
		4	77	77	20.0%	20.0%
		5	139	139	20.0%	20.0%
		6	120	120	20.0%	20.0%
	10	1	149	149	20.0%	20.1%
		2	117	117	20.0%	20.0%
		3	116	116	20.0%	20.0%
		4	69	69	20.0%	19.9%
	11	1	65	65	20.0%	19.9%
		2	72	72	20.0%	19.9%
		3	109	109	20.0%	20.0%
		4	140	140	20.0%	20.0%
		5	42	42	20.0%	19.8%
		6	142	142	20.0%	20.0%
	12	1	145	145	20.0%	20.1%
		2	69	69	20.0%	20.1%
		3	98	98	20.0%	20.1%
		4	134	134	20.0%	20.0%
		5	114	114	20.0%	20.0%
		6	137	137	20.0%	19.9%
中部	2	1	119	119	20.0%	20.1%
		2	153	153	20.0%	19.9%
		3	101	101	20.0%	20.0%
		4	52	52	20.0%	19.8%
		5	144	144	20.0%	20.0%

规划文件: c:\poll.csplan

通过此摘要表（此处显示的是表的顶部）可以复查抽样的第二阶段。该表对于检查抽样是否按计划进行也非常有用。按照要求，从第一阶段抽取的每个镇的每个区抽样大约 20% 的选民。

样本结果

图片 13-46
带有样本结果的数据编辑器



	voteid	nbrhood	town	county	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	SampleWeight_Final_
376	368	4	9	1	0.44	2.26	0.20	11.28	11.28
377	369	4	9	1	0.44	2.26	0.20	11.28	11.28
378	374	4	9	1	0.44	2.26	0.20	11.28	11.28
379	376	4	9	1	0.44	2.26	0.20	11.28	11.28
380	379	4	9	1	0.44	2.26	0.20	11.28	11.28
381	380	4	9	1	0.44	2.26	0.20	11.28	11.28
382	382	4	9	1	0.44	2.26	0.20	11.28	11.28
383	13	5	9	1	0.44	2.26	0.20	11.26	11.26
384	18	5	9	1	0.44	2.26	0.20	11.26	11.26
385	23	5	9	1	0.44	2.26	0.20	11.26	11.26
386	38	5	9	1	0.44	2.26	0.20	11.26	11.26

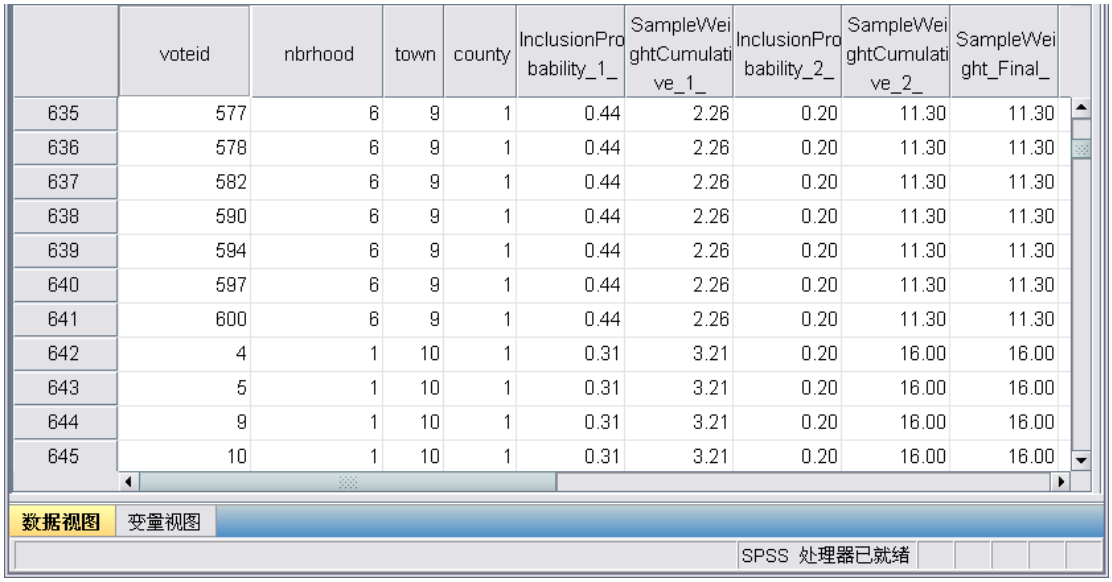
数据视图 变量视图

SPSS 处理器已就绪

可在新创建的数据集中看到抽样结果。 五个新变量保存到了工作文件中，分别表示每个阶段的包含概率和累积抽样权重，以及最终抽样权重。未选入样本的选民排除在此数据集之外。

同一区中，选民的最终抽样权重相等，这是因为选民是根据简单随机抽样方法在区中选择的。但是，同一镇的区之间的最终抽样权重不同，原因在于不是所有区的抽样比例都正好是 20%。

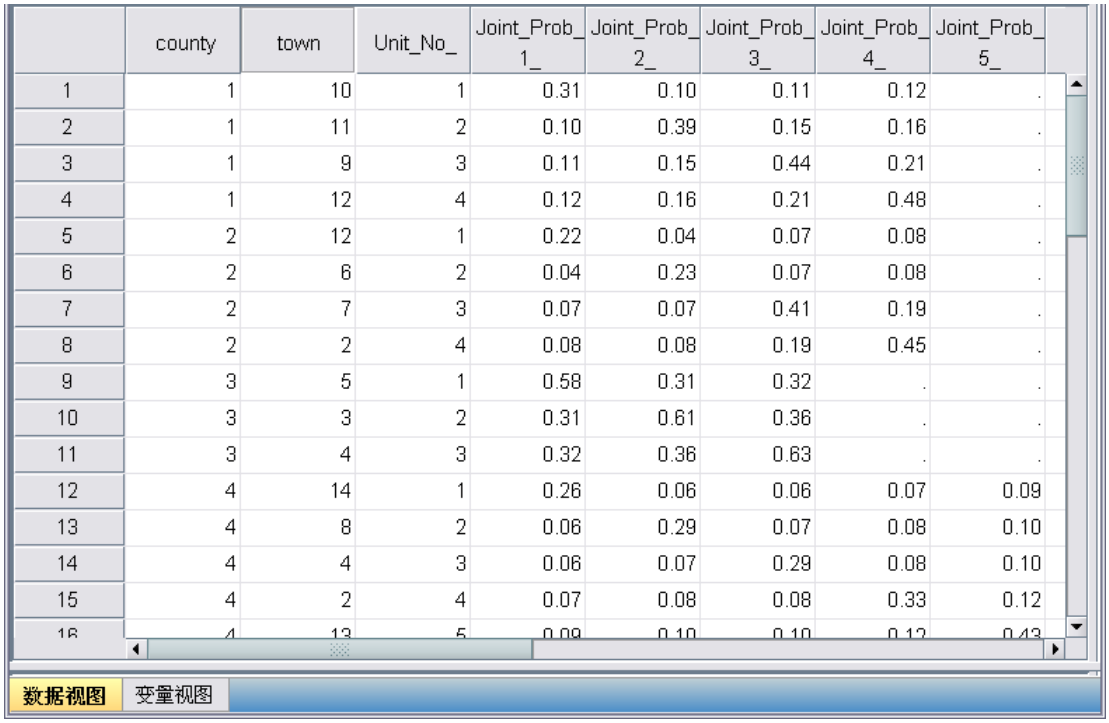
图片 13-47
带有样本结果的数据编辑器



	voteid	nbrhood	town	county	InclusionProbability_1_	SampleWeightCumulative_1_	InclusionProbability_2_	SampleWeightCumulative_2_	SampleWeight_Final_
635	577	6	9	1	0.44	2.26	0.20	11.30	11.30
636	578	6	9	1	0.44	2.26	0.20	11.30	11.30
637	582	6	9	1	0.44	2.26	0.20	11.30	11.30
638	590	6	9	1	0.44	2.26	0.20	11.30	11.30
639	594	6	9	1	0.44	2.26	0.20	11.30	11.30
640	597	6	9	1	0.44	2.26	0.20	11.30	11.30
641	600	6	9	1	0.44	2.26	0.20	11.30	11.30
642	4	1	10	1	0.31	3.21	0.20	16.00	16.00
643	5	1	10	1	0.31	3.21	0.20	16.00	16.00
644	9	1	10	1	0.31	3.21	0.20	16.00	16.00
645	10	1	10	1	0.31	3.21	0.20	16.00	16.00

与第二阶段的选民不同，同一县中的镇的第一阶段抽样权重不相同，原因在于选择概率与大小成比例。

图片 13-48
联合概率文件



	county	town	Unit_No_	Joint_Prob_1_	Joint_Prob_2_	Joint_Prob_3_	Joint_Prob_4_	Joint_Prob_5_
1	1	10	1	0.31	0.10	0.11	0.12	.
2	1	11	2	0.10	0.39	0.15	0.16	.
3	1	9	3	0.11	0.15	0.44	0.21	.
4	1	12	4	0.12	0.16	0.21	0.48	.
5	2	12	1	0.22	0.04	0.07	0.08	.
6	2	6	2	0.04	0.23	0.07	0.08	.
7	2	7	3	0.07	0.07	0.41	0.19	.
8	2	2	4	0.08	0.08	0.19	0.45	.
9	3	5	1	0.58	0.31	0.32	.	.
10	3	3	2	0.31	0.61	0.36	.	.
11	3	4	3	0.32	0.36	0.63	.	.
12	4	14	1	0.26	0.06	0.06	0.07	0.09
13	4	8	2	0.06	0.29	0.07	0.08	0.10
14	4	4	3	0.06	0.07	0.29	0.08	0.10
15	4	2	4	0.07	0.08	0.08	0.33	0.12
16	4	13	5	0.09	0.10	0.10	0.12	0.43

文件 poll_jointprob.sav 包含郡内选定镇区的第一阶段联合概率。郡为第一阶段分层变量，镇区为聚类变量。这些变量的组合唯一地标识所有第一阶段 PSU。单位编号标记每层内的 PSU，用于配合联合概率1、联合概率2、联合概率3、联合概率4 和联合概率5。前两层每层有 4 个 PSU；因此，对于这些层，联合包含概率矩阵为 4×4，并且这些行的联合概率5 列保留为空。类似地，层 3 和 5 具有 3×3 联合包含概率矩阵，层 4 具有 5×5 联合包含概率矩阵。

通过研究联合包含概率矩阵的值可以看出对联合概率文件的需要。当抽样方法不是 PPS WOR 方法时，PSU 的选择相互独立，并且它们的联合包含概率就是它们各自的包含概率之积。而郡 1 的镇区 9 和 10 的联合包含概率约为 0.11（请参见联合概率3 的第一个个案或联合概率1 的第三个个案），或小于它们的各包含概率的乘积（联合概率1 的第一个个案和联合概率3 的第三个个案的乘积为 0.31×0.44=0.1364）。

现在，民意测验专家将访问所选样本。一旦获得结果，即可使用抽样计划 poll.csplan 提供抽样指定项，使用 poll_jointprob.sav 提供需要的联合包含概率，通过复杂样本分析过程对样本进行处理。

相关过程

“复杂样本抽样向导”过程是很有用的工具，可以创建抽样计划文件和抽取样本。

- 要准备样本用于分析，如果不能访问抽样计划文件，则请使用分析准备向导。

章
14

复杂样本分析准备向导

分析准备向导将引导您完成创建或修改分析计划的各个步骤，以用于各种“复杂样本”分析过程。在不能访问用于抽取样本的抽样计划文件时，该向导非常有用。

使用复杂样本分析准备向导准备 NHIS 公共数据

全美国健康访问调查 (NHIS) 是对美国全体公民的大型人口调查。该调查对美国的具有全国代表性的家庭样本进行了面对面的访问，并获取了每个家庭的成员的健康行为和健康状态人口统计信息和观察数据。

一个 2000 份的调查子集收集在 nhis2000_subset.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件. 使用复杂样本分析准备向导为此数据文件创建一个分析计划，以便由复杂样本分析过程处理数据。

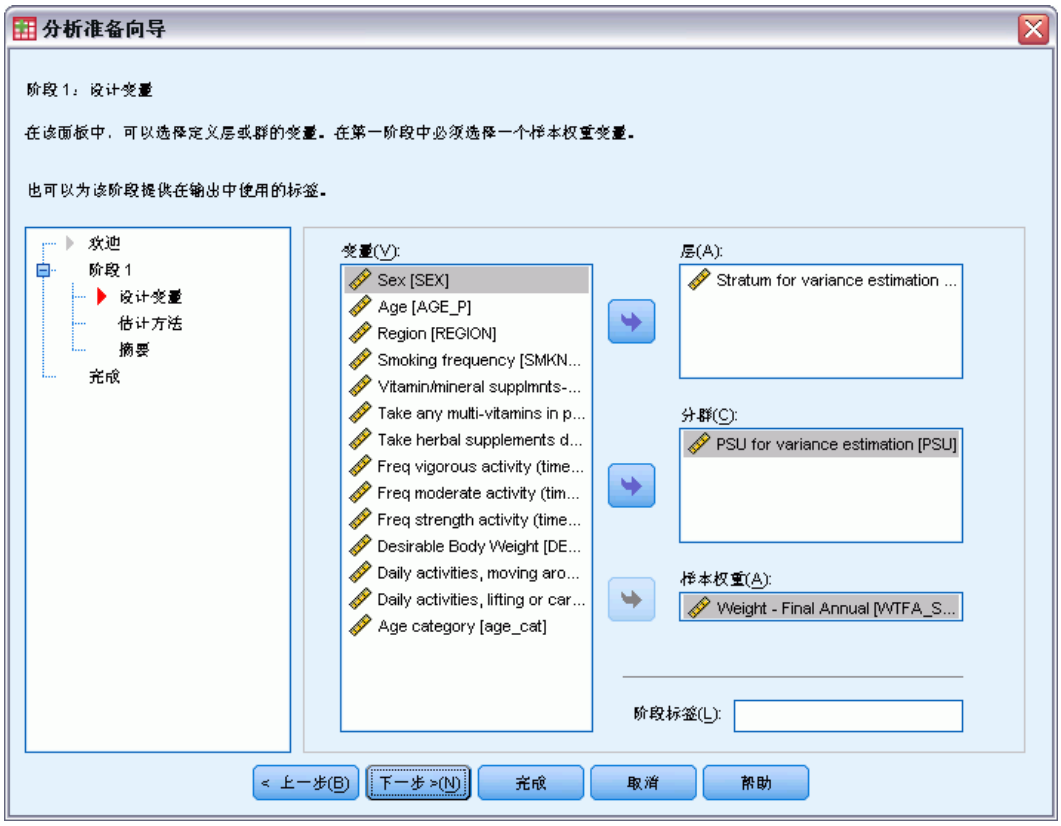
使用向导

- 要使用复杂样本分析准备向导准备一个样本，请从菜单中选择：
  分析 > 复杂抽样 > 准备分析...

图片 14-1
分析准备向导，“欢迎”步骤

- 浏览到您要保存计划文件的位置，输入 nhis2000_subset.csaplan 作为分析计划文件的名称。
- 单击下一步。

图片 14-2
分析准备向导，“设计变量”步骤（阶段 1）



数据是用复杂多阶段样本获取的。但是，对于最终用户，原始 NHIS 设计变量转换为一组简化的设计变量和权重变量，这些变量的结果与原始设计结构的结果近似。

- 选择 Stratum for variance estimation 作为层次变量。
- 选择 PSU for variance estimation 作为聚类变量。
- 选择 Weight - Final Annual 作为样本权重变量。
- 单击**完成**。

摘要

图片 14-3
摘要

			阶段 1
设计变量	分层	1	Stratum for variance estimation
	群集	1	PSU for variance estimation
分析信息	估计量假设		有替换抽样

规划文件:c:
his2000_subset.csplan
权重变量:Weight - Final Annual
SRS 估计量:无替换抽样

通过该摘要表可以复查您的分析计划。计划由一个阶段构成，其设计具有一个分层变量和一个聚类变量。使用放回式 (WR) 估计，计划保存在 c:\nhis2000_subset.csaplan 中。现在通过复杂样本分析过程，可以使用此计划文件处理 nhis2000_subset.sav。

抽样权重不在数据文件中时准备分析

信贷员有一组客户记录，这些记录是根据一项复杂设计获得的；但是文件中不包含抽样权重。该信息包含在 bankloan_cs_noweights.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件. 以对抽样设计的了解为起点，信贷员希望使用复杂样本分析准备向导为此数据文件创建分析计划，以便通过复杂样本分析过程进行处理。

信贷员知道记录是在两个阶段选择的，在第一阶段中，以相等概率从 100 个银行分支机构中以不放回方式选择了 15 个。然后在第二阶段中，以相等概率从选出的每个分支机构中采用不放回方式选择 100 个客户，每个分支机构的客户数量信息都包括在数据文件中。创建分析计划的第一步是计算分阶段包含概率和最终抽样权重。

计算包含概率和抽样权重

- 要计算第一阶段的包含概率，请从菜单中选择：
  转换 > **计算变量...**

图片 14-4
“计算变量”对话框

100 个银行分支机构中的 15 个是在第一阶段以不放回方式选择的；因此，一个给定分支机构的选择概率为 15/100 = 0.15。

- 键入 inclprob_s1 作为目标变量。
- 键入 0.15 作为数值表达式。
- 单击确定。

图片 14-5
“计算变量”对话框

100 个客户是在第二阶段从每个分支机构选择的；因此，某个给定分支机构的一个给定客户的阶段 2 包含概率为 100/该分支机构的客户数。

- 调用“计算变量”对话框。
- 键入 inclprob_s2 作为目标变量。
- 键入 100/ncust 作为数值表达式。
- 单击确定。

图片 14-6
“计算变量”对话框

现在，每个阶段都有了包含概率，可以很方便地计算最终抽样权重。

- 调用“计算变量”对话框。
- 键入 finalweight 作为目标变量。
- 键入 1/(inclprob_s1 * inclprob_s2) 作为数值表达式。
- 单击确定。

  现在即可创建分析计划。

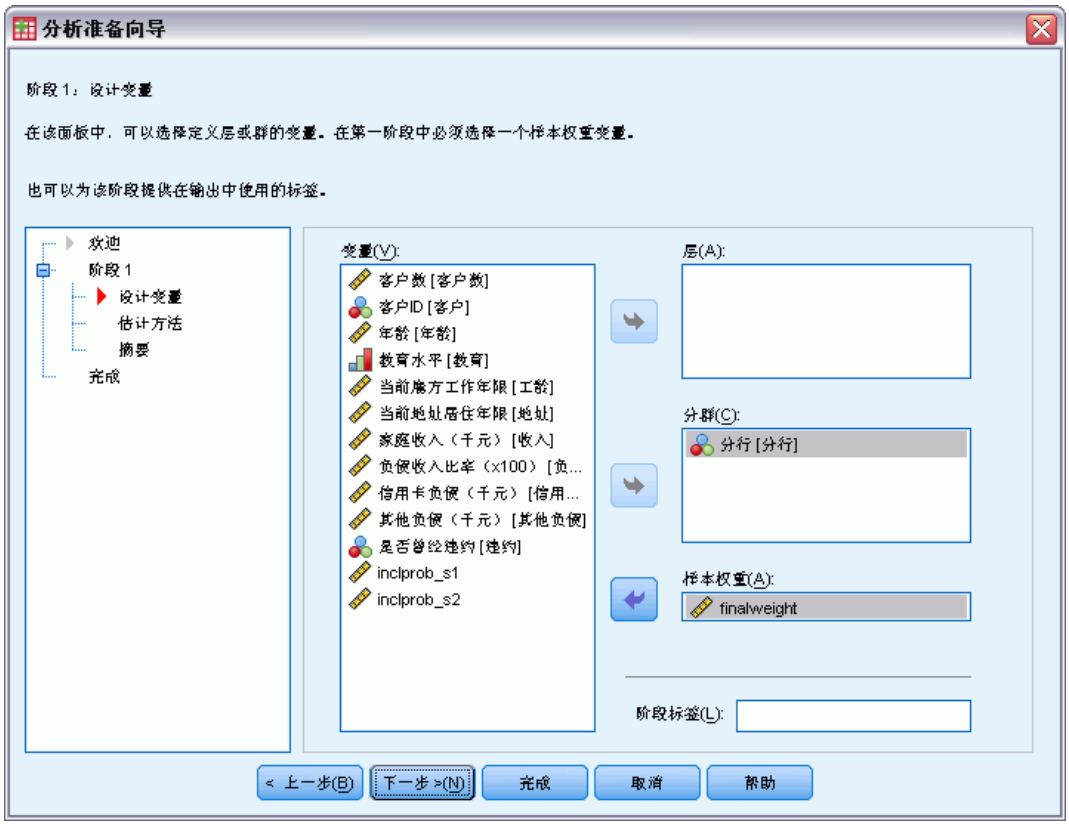
使用向导

- 要使用复杂样本分析准备向导准备一个样本，请从菜单中选择：
  分析 > 复杂抽样 > 准备分析...

图片 14-7
分析准备向导，“欢迎”步骤

- 浏览到您要保存计划文件的位置，输入 bankloan.csaplan 作为分析计划文件的名称。
- 单击下一步。

图片 14-8
分析准备向导，“设计变量”步骤（阶段 1）



- 选择 Branch 作为聚类变量。
- 选择 finalweight 作为样本权重变量。
- 单击下一步。

图片 14-9
分析准备向导，“估计方法”步骤（阶段 1）

- 选择等概率 WOR 作为第一阶段估计方法。
- 单击下一步。

图片 14-10
分析准备向导，“大小”步骤（阶段 1）

- 选择**从变量中读取值**并选择 inclprob_s1 作为包含第一阶段包含概率的变量。
- 单击**下一步**。

图片 14-11
分析准备向导，“计划摘要”步骤（阶段 1）

- 选择是，现在添加阶段 2。
- 单击下一步，然后在“设计变量”步骤中单击下一步。

图片 14-12
分析准备向导，“估计方法”步骤（阶段 2）

- 选择等概率 WOR 作为第二阶段估计方法。
- 单击下一步。

图片 14-13
分析准备向导，“大小”步骤（阶段 2）

- 选择从变量中读取值并选择 inclprob_s2 作为包含第二阶段包含概率的变量。
- 单击完成。

摘要

图片 14-14
摘要表

			阶段 1	阶段 2
设计变量	群集	1	Branch	
分析信息	估计量假设		无替换等概率抽样	无替换等概率抽样
	包含概率		从变量 inclprob_s1 获得	从变量 inclprob_s2 获得

规划文件: c:\bankloan.csplan
权重变量: finalweight
SRS 估计量:无替换抽样

通过该摘要表可以复查您的分析计划。计划由两个阶段构成，其设计具有一个聚类变量。使用等概率不放回方式 (WOR) 估计，计划保存在 c:\bankloan.csaplan 中。现在，通过复杂样本分析过程，即可使用此计划文件处理 bankloan_noweights.sav（具有已计算的包含概率和抽样权重）。

相关过程

“复杂样本分析准备向导”过程是一个有用的工具，可以在无法访问抽样计划文件时准备分析样本。

- 要创建抽样计划文件并抽取样本，请使用抽样向导。

章
15

复杂样本频率

“复杂样本频率”过程可以为所选变量生成频率表并显示单变量统计。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

使用复杂样本频率分析营养补充品的使用情况

研究人员希望使用全美国健康访问调查 (NHIS) 的结果和以前创建的分析计划对美国公民的营养补充品使用情况进行研究。有关详细信息，请参阅第 133 页码第 14 章中的使用复杂样本分析准备向导准备 NHIS 公共数据.

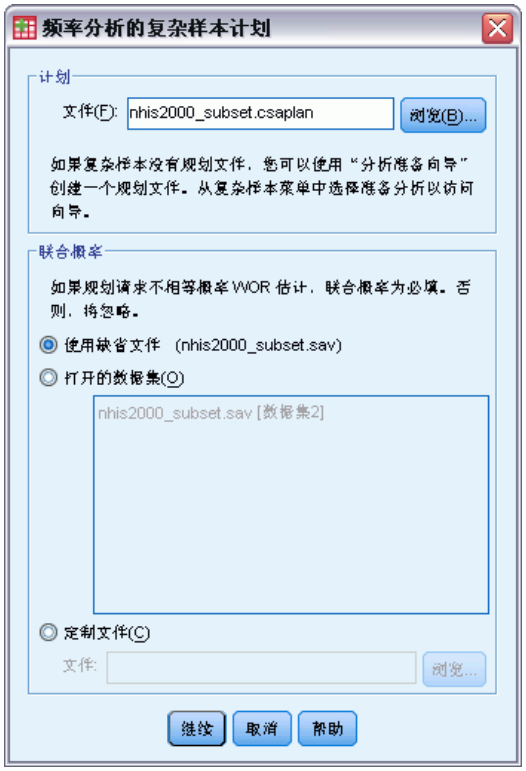
一个 2000 份的调查子集收集在 nhis2000_subset.sav 中。分析计划存储在 nhis2000_subset.csaplan中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.

使用复杂样本频率生成营养补充品使用情况的统计量。

运行分析

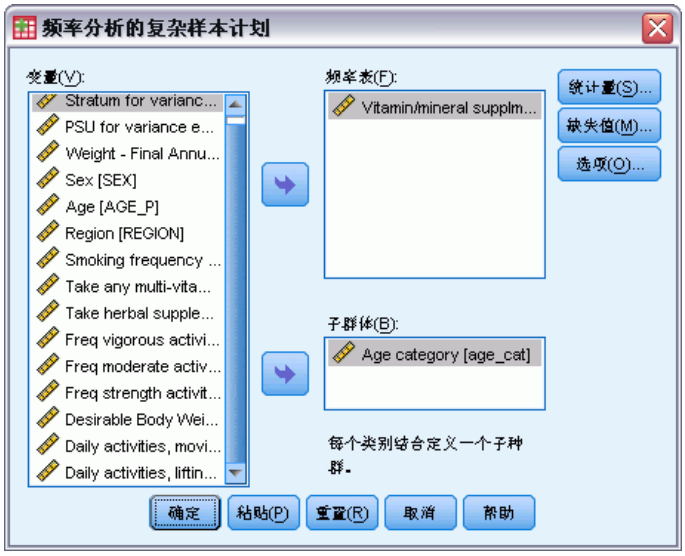
- 要运行复杂样本频率分析，请从菜单中选择：
  分析 > **复杂样本** > **频率...**

图片 15-1
“复杂样本计划”对话框



- 浏览至 nhis2000_subset.csaplan 并将其选中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.
- 单击继续。

图片 15-2
“频率”对话框



- 选择 Vitamin/mineral supplmnts-past 12 m 作为频率变量。
- 选择 Age category 作为一个子体变量。
- 单击**统计量**。

图片 15-3
“频率：统计量”对话框

- 选择“单元格”组中的**表百分比**。
- 在“统计量”组中选择**置信区间**。
- 单击**继续**。
- 在“频率”对话框中单击**确定**。

频率表

图片 15-4
变量/情况的频率表

		估计	标准误差	95% 置信区间	
				下限	下限
种群大小	Yes	102767095	1185126.7	100435967	105098223
	No	90794234	1094401.9	88641560	92946908
	合计	193561329	1789098.7	190042196	197080462
合计 %	Yes	53.1%	.4%	52.4%	53.8%
	No	46.9%	.4%	46.2%	47.6%
	合计	100.0%	.0%	100.0%	100.0%

对选择的每个单元格度量计算选择的每个统计量。第一列包含使用或不使用维生素/矿物质补充品的人口数量和百分比的估计值。 置信区间不重叠；因此，可以得出结论，在整体上，使用维生素/矿物质补充品的美国人比不使用的美国人多。

基于子体的频率

图片 15-5
基于子体的频率表

Age category			估计	标准误差	95% 置信区间	
					下限	下限
18-24	种群大小	Yes	10018312	350602.35	9328681.9	10707942
		No	15472368	499182.39	14490483	16454253
		合计	25490680	680732.81	24151688	26829672
	合计 %	Yes	39.3%	1.0%	37.4%	41.2%
		No	60.7%	1.0%	58.8%	62.6%
		合计	100.0%	.0%	100.0%	100.0%
25-44	种群大小	Yes	39163840	660855.72	37863946	40463734
		No	39503150	645934.19	38232606	40773694
		合计	78666990	961114.33	76776491	80557489
	合计 %	Yes	49.8%	.6%	48.7%	50.9%
		No	50.2%	.6%	49.1%	51.3%
		合计	100.0%	.0%	100.0%	100.0%
45-64	种群大小	Yes	34154952	598603.73	32977507	35332397
		No	24005512	497723.83	23026496	24984528
		合计	58160464	814680.41	56557999	59762929
	合计 %	Yes	58.7%	.6%	57.5%	60.0%
		No	41.3%	.6%	40.0%	42.5%
		合计	100.0%	.0%	100.0%	100.0%
65+	种群大小	Yes	19429991	439459.79	18565580	20294402
		No	11813204	314238.08	11195102	12431306
		合计	31243195	587623.44	30087348	32399042
	合计 %	Yes	62.2%	.7%	60.7%	63.6%
		No	37.8%	.7%	36.4%	39.3%
		合计	100.0%	.0%	100.0%	100.0%

计算基于子体的统计量时，会基于 Age category 的值计算所选每个单元格度量的每个选定统计量。第一列包含每个类别使用或不使用维生素/矿物质补充品的人口数量和百分比的估计值。 表百分比的置信区间不重叠；因此，可以得出结论，随着年龄的增大，维生素/矿物质补充品的使用者在增加。

摘要

使用“复杂样本频率”过程，已获得美国公民营养补充品使用情况的统计量。

- 整体上，使用维生素/矿物质补充品的美国人比不使用的美国人多。
- 如果按年龄组分隔，使用维生素/矿物质补充品的美国人比例随年龄的增加而增大。

相关过程

“复杂样本频率”过程是很有用的工具，可以获取观察数据（通过复杂抽样设计获取）的分类变量的单变量描述统计量。

- 复杂样本抽样向导用于指定复杂抽样设计指定项并获取一个样本。抽样向导创建的抽样计划文件包含一个缺省的分析计划，当您分析据此计划获取的样本时，可以在“计划”对话框中指定该抽样计划文件。
- 复杂样本分析准备向导用于为现有的复杂样本设置分析指定项。当您分析与该计划对应的样本时，可以在“计划”对话框中指定由抽样向导创建的分析计划文件。
- 复杂样本交叉表过程提供分类变量交叉制表的描述统计量。
- 复杂样本描述过程提供刻度变量的单变量描述统计量。

章
16

复杂样本描述

“复杂样本描述”过程为多个变量显示单变量摘要统计量。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

使用复杂样本描述分析活动水平

研究人员希望使用全美国健康访问调查 (NHIS) 的结果和以前创建的分析计划对美国公民的活动水平进行研究。有关详细信息，请参阅第 133 页码第 14 章中的使用复杂样本分析准备向导准备 NHIS 公共数据.

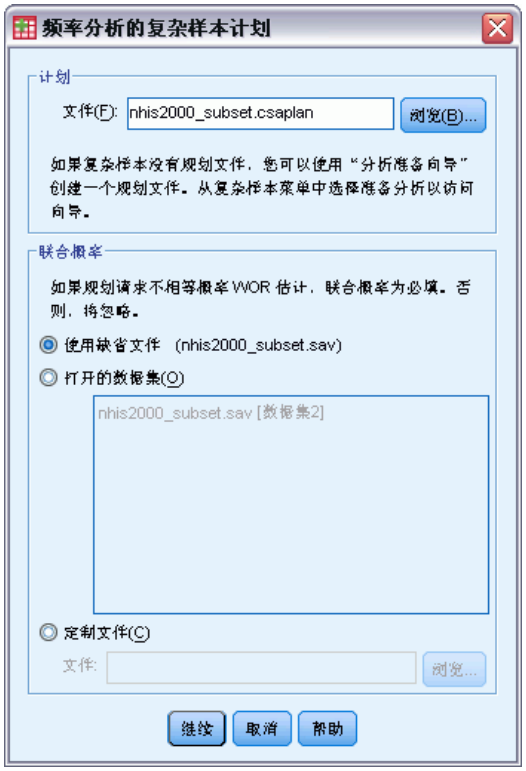
一个 2000 份的调查子集收集在 nhis2000_subset.sav 中。分析计划存储在 nhis2000_subset.csaplan中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.

使用复杂样本描述生成活动水平的单变量描述统计量。

运行分析

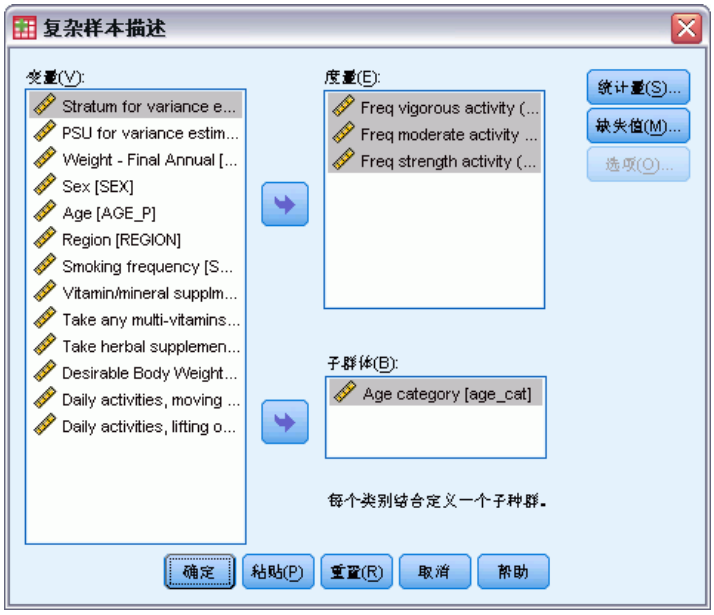
- 要运行复杂样本描述分析，请从菜单中选择：
  分析 > **复杂样本** > **描述**...

图片 16-1
“复杂样本计划”对话框



- 浏览至 nhis2000_subset.csaplan 并将其选中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.
- 单击继续。

图片 16-2
“描述”对话框



- 选择 Freq vigorous activity (times per wk) 到 Freq strength activity (times per wk) 作为测量变量。
- 选择 Age category 作为一个子体变量。
- 单击统计量。

图片 16-3
“描述：统计量”对话框

- 在“统计量”组中选择置信区间。

- ▶ 单击**继续**。
- ▶ 在“复杂样本描述”对话框中单击**确定**。

单变量统计

图片 16-4
单变量统计

		估计	标准误差	95% 置信区间	
				下限	下限
均值	Freq vigorous activity (times per wk)	3.73	.033	3.66	3.79
	Freq moderate activity (times per wk)	4.90	.041	4.82	4.98
	Freq strength activity (times per wk)	3.52	.042	3.43	3.60

对每个测量变量计算所选的每个统计量。第一列包含个人每周参与特定活动类型的平均时间量的估计值。 均值的置信区间不重叠。因此，可以得出结论，在整体上，美国人参与强度活动的时间比参与剧烈活动少，而参与剧烈活动的时间比参与中度活动的时间少。

基于子体的单变量统计

图片 16-5
基于子体的单变量统计

Age category			估计	标准误差	95% 置信区间	
					下限	下限
18-24	均值	Freq vigorous activity (times per wk)	3.92	.087	3.75	4.09
		Freq moderate activity (times per wk)	5.18	.137	4.91	5.45
		Freq strength activity (times per wk)	3.45	.085	3.28	3.62
25-44	均值	Freq vigorous activity (times per wk)	3.55	.048	3.46	3.65
		Freq moderate activity (times per wk)	4.73	.056	4.62	4.84
		Freq strength activity (times per wk)	3.28	.052	3.18	3.38
45-64	均值	Freq vigorous activity (times per wk)	3.79	.063	3.66	3.91
		Freq moderate activity (times per wk)	4.88	.070	4.74	5.02
		Freq strength activity (times per wk)	3.65	.092	3.47	3.84
65+	均值	Freq vigorous activity (times per wk)	4.18	.111	3.96	4.39
		Freq moderate activity (times per wk)	5.22	.084	5.06	5.39
		Freq strength activity (times per wk)	4.66	.155	4.36	4.97

按 Age category 的值计算每个测量变量的每个所选统计量。第一列包含每个类别的个人每周参与特定活动类型的平均时间量的估计值。 使用均值的置信区间，可以得出一些有趣的结论。

- 对于剧烈活动和中度活动，25 - 44 岁的活动时间比 18 - 24 和 45 - 64 岁的少，45 - 64 岁的活动时间比 65 岁或 65 岁以上的少。
- 对于强度活动，25 - 44 岁的活动时间比 45 - 64 岁的少，18 - 24 和 45 - 64 岁的活动时间比 65 岁或 65 岁以上的少。

摘要

使用“复杂样本描述”过程，已获得美国公民活动水平的统计量。

- 整体上，美国人在不同活动类型上所花的时间量是不同的。
- 按照年龄分隔，大体情况是，美国人刚大学毕业时，活动时间比在校时少，但随着年龄增加，会更勤于运动。

相关过程

“复杂样本描述”过程是很有用的工具，可以获取观察数据（通过复杂抽样设计获取）的刻度度量的单变量描述统计量。

- 复杂样本抽样向导用于指定复杂抽样设计指定项并获取一个样本。抽样向导创建的抽样计划文件包含一个缺省的分析计划，当您分析据此计划获取的样本时，可以在“计划”对话框中指定该抽样计划文件。
- 复杂样本分析准备向导用于为现有的复杂样本设置分析指定项。当您分析与该计划对应的样本时，可以在“计划”对话框中指定由抽样向导创建的分析计划文件。
- 复杂样本比率过程提供刻度度量比率的描述统计量。
- 复杂样本频率过程提供分类变量的单变量描述统计量。

章
17

复杂样本交叉表

复杂样本交叉表过程可以为所选变量对生成交叉表并显示二阶统计量。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

使用复杂样本交叉表度量事件的相对风险

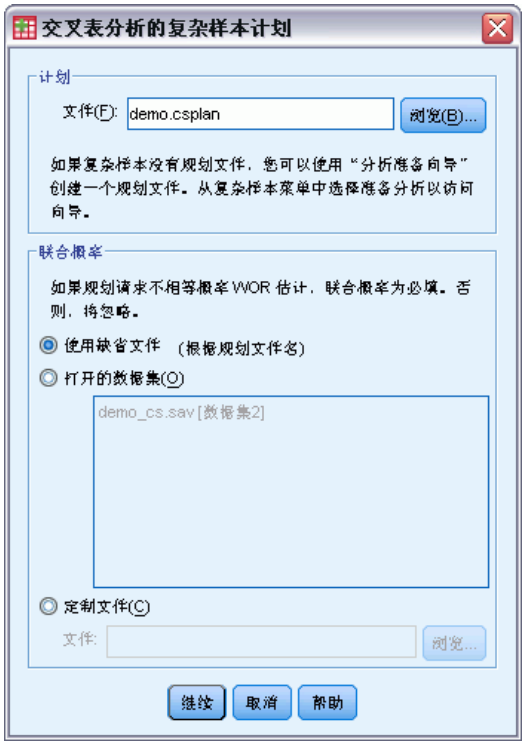
一个以传统方式销售杂志订阅的公司每月按购买的姓名数据库发送邮件。响应率通常很低，因此，需要找到一种方法更好地确定潜在客户。一个建议是，根据阅读报纸的人更可能订阅杂志的假设，集中向报纸订户发送邮件。

通过构造 Newspaper subscription 和 Response 的 2x2 表，并计算报纸订户响应邮件的相对风险，使用“复杂样本交叉表”过程检验这一理论。该信息收集在 demo_cs.sav 中并且应该使用抽样计划文件 demo.csplan 来对其进行分析。 有关详细信息，请参阅第 251 页码附录 A 中的样本文件.

运行分析

▶ 要运行复杂样本交叉表分析，请从菜单中选择：
分析 > 复杂样本 > 交叉表...

图片 17-1
“复杂样本计划”对话框

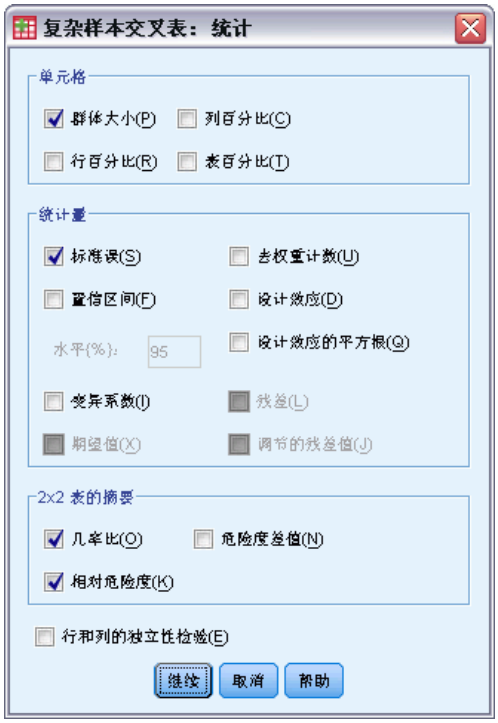


- 浏览至 demo.csplan 并将其选中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.
- 单击继续。

图片 17-2
“交叉表”对话框

- 选择 Newspaper subscription 作为行变量。
- 选择 Response 作为列变量。
- 查看按收入类别分隔的结果也有一定意义，因此选择 Income category in thousands 作为子体变量。
- 单击**统计量**。

图片 17-3
“交叉表：统计量”对话框



- 取消选择**群体大小**，在“单元”组中选择**行百分比**。
- 在“2x2 表的摘要”组中选择**几率比**和**相对危险度**。
- 单击**继续**。
- 在“复杂样本交叉表”对话框中单击**确定**。

通过这些选择，会为 Newspaper subscription 和 Response 生成交叉表和风险估计。还会创建按 Income category in thousands 分隔的不同表。

交叉制表

图片 17-4
基于响应的报纸订阅交叉表

Newspaper subscription			Response		
			Yes	No	合计
Yes	在 Newspaper subscription 中的 %	估计	18.6%	81.4%	100.0%
		标准误差	1.2%	1.2%	.0%
No	在 Newspaper subscription 中的 %	估计	10.6%	89.4%	100.0%
		标准误差	.7%	.7%	.0%
合计	在 Newspaper subscription 中的 %	估计	13.5%	86.5%	100.0%
		标准误差	.7%	.7%	.0%

交叉制表显示整体上很少人响应邮件。但是，报纸订户的响应比例更大些。

风险估计

图片 17-5
基于响应的报纸订阅风险估计

			估计
Newspaper subscription * Response	几率比		1.912
	相对风险	适用于群组 Response = Yes	1.743
		适用于群组 Response = No	.911

仅可计算全部单元格都含观测值的 2x2 表的统计量。

相对风险是事件概率的比率。邮件响应的相对风险是报纸订户响应的概率与非订户响应的概率的比率。因此，相对风险的估计值为 17.2%/10.3% = 1.673。 同样，无响应的相对风险是订户不响应的概率与非订户不响应的概率的比率。此相对风险的估计值为 0.923。根据这些结果，可以估计，报纸订户响应邮件的可能性是非订户的 1.673 倍，或不响应邮件的可能性是非订户的 0.923 倍。

几率比是事件几率的比率。事件的几率是事件发生的概率与事件不发生的概率的比率。因此，报纸订户响应邮件的几率估计值为 17.2%/82.8% = 0.208。 同样，非订户响应的几率估计值为 10.3%/89.7% = 0.115。因此，几率比的估计值为 0.208/0.115 = 1.812（注意干涉步骤中存在一定的舍入误差）。几率比也是响应的相对风险与不响应的相对风险的比率，即 1.673/0.923 = 1.812。

几率比与相对风险

由于是比率的比率，几率比很难解释。相对风险的则较容易解释，因此仅有几率比并不很有帮助。但是，某些常见情况下，相对风险的估计值不理想，而几率比可用作近似事件的相对风险。满足以下两种条件时，几率比应该用作事件的相对风险的近似值：

- 关注事件的概率很小 (< 0.1)。这个条件确保几率比能很好地近似相对风险。在本示例中，事件是对邮件的响应。
- 研究所用的设计是个案控制的。这个条件表示通常的相对风险估计值不会很理想。个案控制研究是回顾性的，常用于关注事件发生的可能性很小，或者预期试验的设计不实际或不道德的情况。

本示例中，这两个条件都不满足，响应者的整体比例为 12.8%，研究的设计也不是个案控制的，因此，将 1.673（而不是几率比的值）作为相对风险报告较为安全。

基于子体的风险估计

图片 17-6
基于响应的报纸订阅风险估计，按收入类别控制

Income category				估计
Under $25	Newspaper subscription * Response	几率比		2.608
		相对风险	适用于群组 Response = Yes	2.149
			适用于群组 Response = No	.824
$25 - $49	Newspaper subscription * Response	几率比		1.923
		相对风险	适用于群组 Response = Yes	1.737
			适用于群组 Response = No	.903
$50 - $74	Newspaper subscription * Response	几率比		1.538
		相对风险	适用于群组 Response = Yes	1.493
			适用于群组 Response = No	.971
$75+	Newspaper subscription * Response	几率比		1.200
		相对风险	适用于群组 Response = Yes	1.182
			适用于群组 Response = No	.985

仅可计算全部单元格都含观测值的 2x2 表的统计量。

分别计算每个收入类别的相对风险估计值。请注意：报纸订户的正响应相对风险随着收入的增加呈逐渐下降的趋势，表示可以进一步确定邮件发送目标。

摘要

使用复杂样本交叉表风险估计值，通过确定报纸订户目标，可以提高对直接发送邮件的响应率。此外，还可以发现一些事实，即不同 Income category 的风险估计值可能不相同，因此，通过确定较低收入的报纸订户目标，可以进一步提高响应率。

相关过程

“复杂样本交叉表”过程是很有用的工具，可以获取观察数据（通过复杂抽样设计获取）的分类变量交叉制表的描述统计量。

- 复杂样本抽样向导用于指定复杂抽样设计指定项并获取一个样本。抽样向导创建的抽样计划文件包含一个缺省的分析计划，当您分析据此计划获取的样本时，可以在“计划”对话框中指定该抽样计划文件。
- 复杂样本分析准备向导用于为现有的复杂样本设置分析指定项。当您分析与该计划对应的样本时，可以在“计划”对话框中指定由抽样向导创建的分析计划文件。
- 复杂样本频率过程提供分类变量的单变量描述统计量。

章
18

复杂样本比率

“复杂样本比率”过程显示变量的比率的单变量摘要统计。您还可以按子组请求统计量，子组由一个或多个分类变量定义。

使用复杂样本比率辅助进行资产价值评估

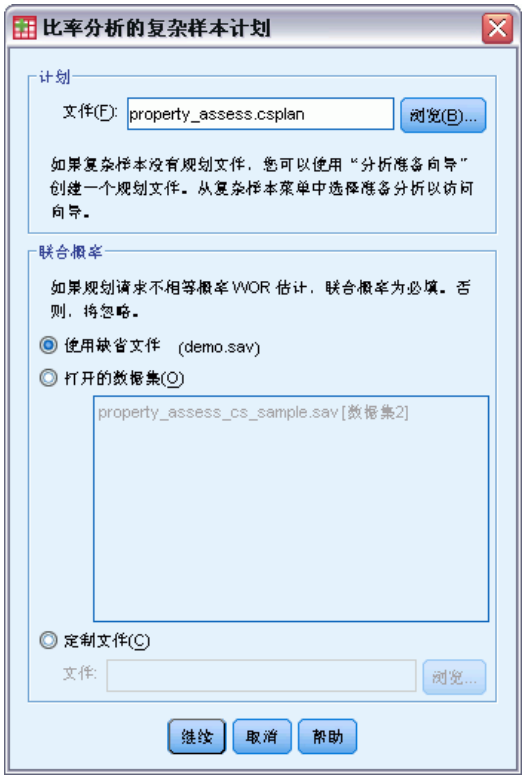
一个州政府机构负责确保对各县的资产税进行公平评估。税是根据资产评估值确定的，因此，该机构希望跟踪各县的资产价值，以确保每个县的记录都是最新的。由于获取当前评估值的资源有限，该机构选择使用复杂样本方法来选择资产。

选择的资产样本及其当前评估信息收集在 property_assess_cs_sample.sav中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.使用复杂样本比率对五个县上次评估以来的资产价值变化进行评估。

运行分析

▶ 要运行复杂样本比率分析，请从菜单中选择：
分析 > 复杂样本 > 比率...

图片 18-1
“复杂样本计划”对话框

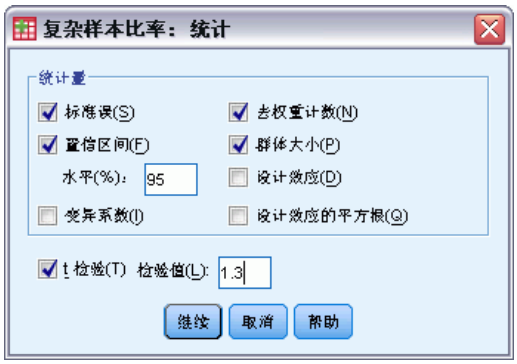


- 浏览至 property_assess.csplan 并将其选中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.
- 单击继续。

图片 18-2
“比率”对话框

- 选择 Current value 作为分子变量。
- 选择 Value at last appraisal 作为分母变量。
- 选择 County 作为子群体变量。
- 单击**统计量**。

图片 18-3
“比率：统计量”对话框



- 在“统计量”组中选择**置信区间**、**去权重计数**和**群体大小**。
- 选择 **t 检验**，输入 1.3 作为检验值。
- 单击**继续**。
- 在“复杂样本比率”对话框中单击**确定**。

比率

图片 18-4
比率表

County	分子	分母	比值估计	标准误差	95% 置信区间		
					下限	下限	检
Eastern	Current value	Value at last appraisal	1.381	.068	1.236	1.525	
Central	Current value	Value at last appraisal	1.364	.064	1.227	1.502	
Western	Current value	Value at last appraisal	1.524	.053	1.410	1.638	
Northern	Current value	Value at last appraisal	1.277	.032	1.208	1.346	
Southern	Current value	Value at last appraisal	1.195	.029	1.134	1.256	

该表的缺省显示较宽，因此，为便于查看，需要对其进行透视。

透视比率表

- ▶ 双击枢轴表以激活它。
- ▶ 从“浏览器”菜单中选择：
  透视 > 透视托盘
- ▶ 依次将 Numerator 和 Denominator 从行拖到层。
- ▶ 将 County 从行拖到列。
- ▶ 将统计量从列拖到行。
- ▶ 关闭透视托盘窗口。

透视比率表

图片 18-5
透视比率表

分母: Value at last appraisal
分子: Current value

统计量		County				
		Eastern	Central	Western	Northern	Southern
比值估计		1.381	1.364	1.524	1.277	1.195
标准误差		.068	.064	.053	.032	.029
95% 置信区间	下限	1.236	1.227	1.410	1.208	1.134
	下限	1.525	1.502	1.638	1.346	1.256
假设检验	检验值	1.3	1.3	1.3	1.3	1.3
	t	1.191	.997	4.201	-.702	-3.646
	df	15	15	15	15	15
	Sig.	.252	.334	.001	.493	.002
种群大小		1883.250	1557.500	4044.000	2306.250	2204.000
未加权计数		168	179	202	205	220

现在比率表已经过透视，可以更方便地比较各县的统计量。

- 比率估计值的范围最低为南县的 1.195，最高为西县的 1.524。
- 标准误之间也有相当差异，最低为南县的 0.029，最高为东县的 0.068。

- 部分置信区间不重叠；因此，可以得出结论，西县的比率高于北县和南县的比率。
- 最后，为了更客观的测量，注意到西县和南县的 t 检验的显著性值小于 0.05。因此，可以得出结论，西县的比率大于 1.3，南县的比率小于 1.3。

摘要

使用“复杂样本比率”过程，已获得 Current value 与 Value at last appraisal 的比率的各种统计量。结果显示，各县资产税的评估中存在一定程度的不平衡，即：

- 西县的比率较高，表明其资产价值评估的记录没有其他县的记录新。该县的资产税可能太低。
- 南县的比率低，表明其资产价值评估的记录比其他县的记录新。该县的资产税可能太高。
- 南县比西县的比率低，但仍然在客观目标 1.3 范围内。

用于跟踪南县资产价值的资源将重新分配给西县，使这些县的比率与其他县一致，并与目标 1.3 一致。

相关过程

“复杂样本比率”过程是很有用的工具，可以获取观察数据（通过复杂抽样设计获取）的刻度度量比率的单变量描述统计量。

- 复杂样本抽样向导用于指定复杂抽样设计指定项并获取一个样本。抽样向导创建的抽样计划文件包含一个缺省的分析计划，当您分析据此计划获取的样本时，可以在“计划”对话框中指定该抽样计划文件。
- 复杂样本分析准备向导用于为现有的复杂样本设置分析指定项。当您分析与该计划对应的样本时，可以在“计划”对话框中指定由抽样向导创建的分析计划文件。
- 复杂样本描述过程提供刻度变量的描述统计量。

章
19

复杂样本一般线性模型

“复杂样本一般线性模型”(CSGLM) 过程对通过复杂抽样方法抽取的样本执行线性回归分析以及方差和协方差分析。您还可以请求对子体进行分析。

使用复杂样本一般线性模型拟合双因子 ANOVA

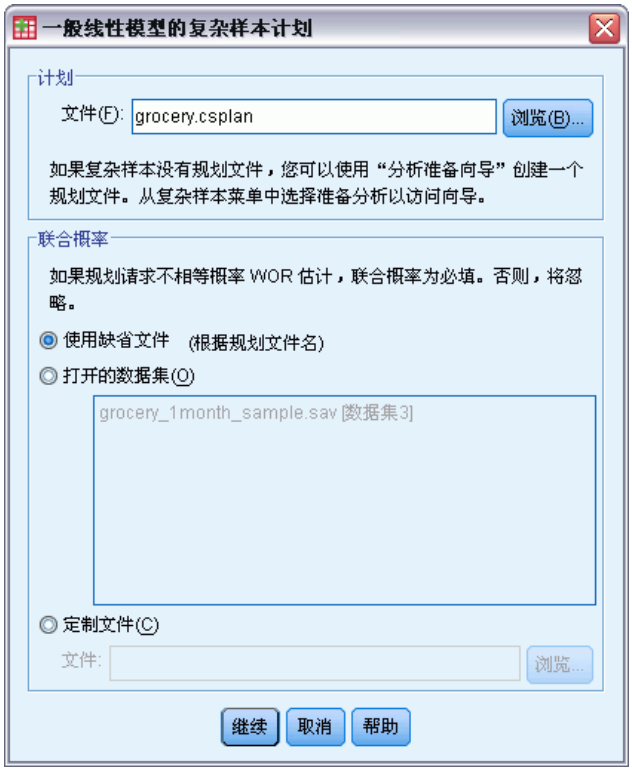
根据一项复杂设计，杂货连锁店对一组顾客的购物习惯进行调查。在获得了调查结果以及每个顾客在上个月的消费金额之后，商店希望了解顾客购物的频率是否与他们在一个月中的消费金额有关，从而针对顾客性别进行控制并采用抽样设计。

该信息收集在 grocery_1month_sample.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.使用“复杂样本一般线性模型”过程对消费金额进行双因子（或双向）ANOVA。

运行分析

► 要运行复杂样本一般线性模型分析，请从菜单中选择：
分析 > 复杂样本 > 一般线性模型...

图片 19-1
“复杂样本计划”对话框

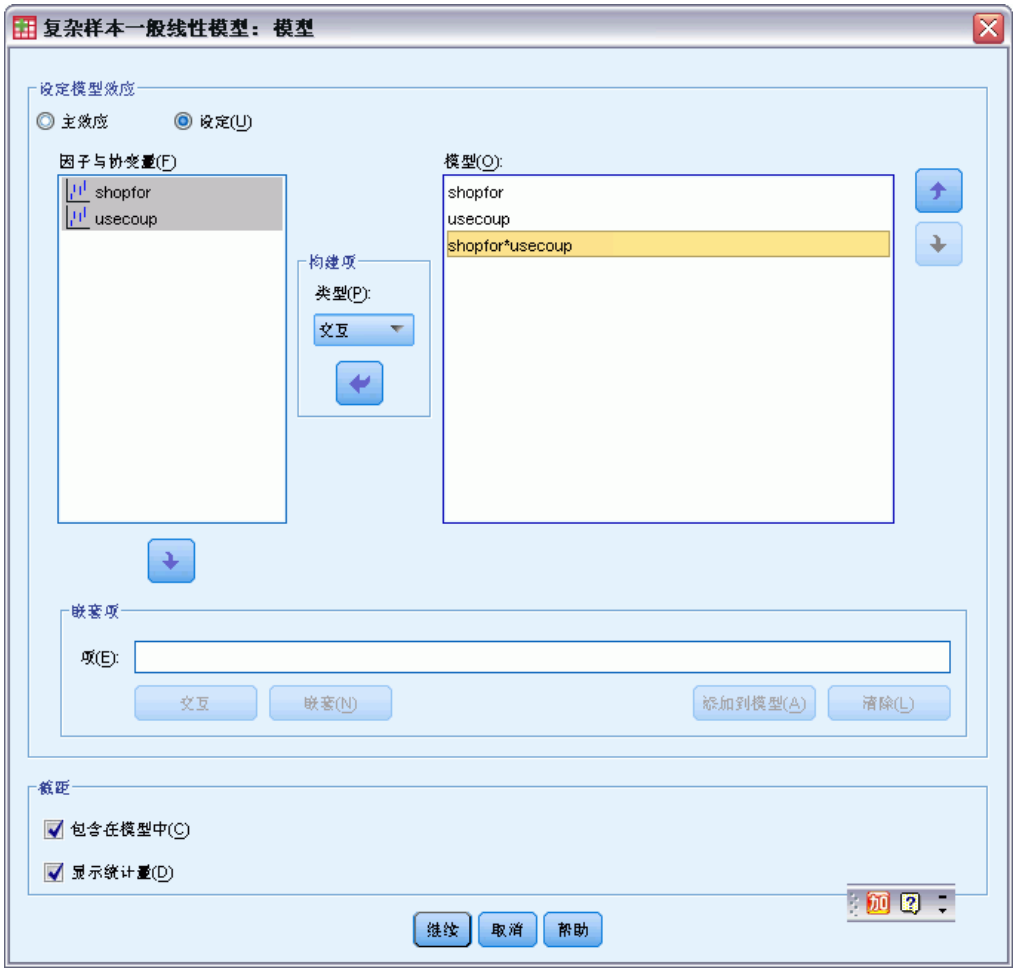


- 浏览至 grocery.csplan 并将其选中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.
- 单击继续。

图片 19-2
“一般线性模型”对话框

- 选择消费金额作为因变量。
- 选择购物者和使用优惠券作为因子。
- 单击**模型**。

图片 19-3
“模型”对话框

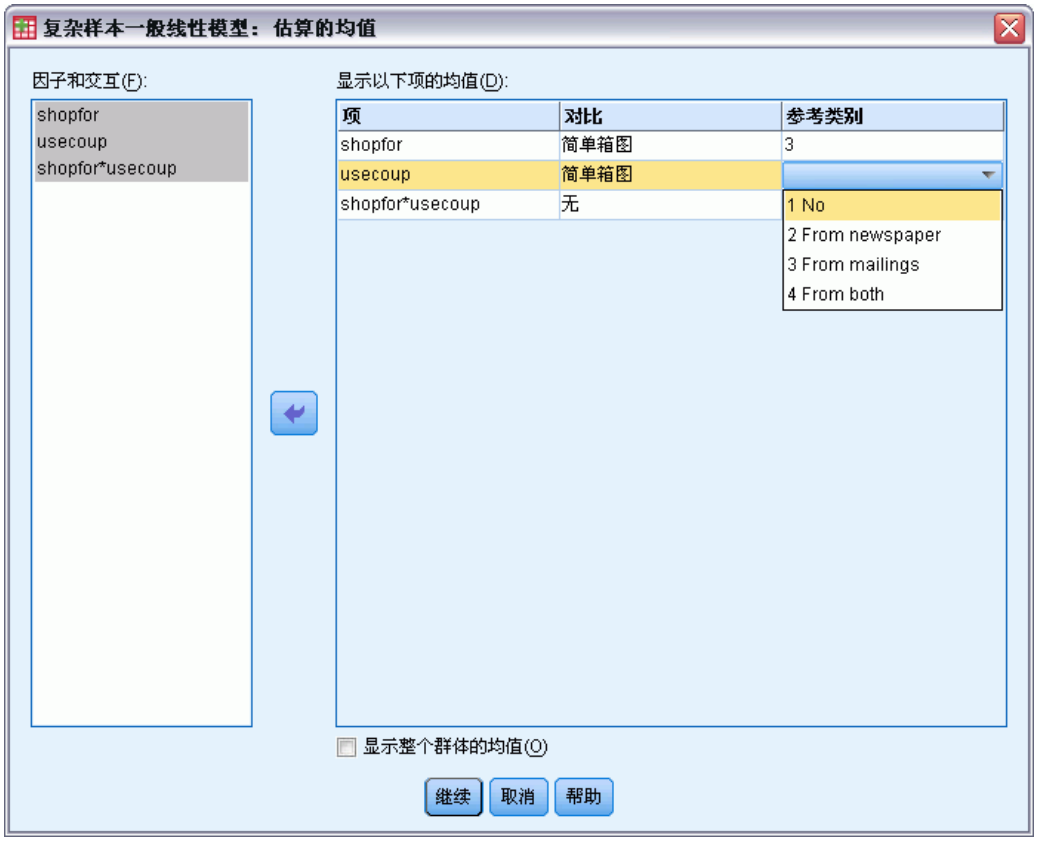


- 选择建立**定制**模型。
- 选择**主效应**作为要建立的项类型，选择 shopfor 和 usecoup 作为模型项。
- 选择**交互**作为要建立的项类型，添加 shopfor*usecoup 交互作为一个模型项。
- 单击**继续**。
- 单击“一般线性模型”对话框中的**统计量**。

图片 19-4
“一般线性模型：统计量”对话框

- 选择“模型参数”组中的**估算**、**标准误**、**置信区间**以及**设计效应**。
- 单击**继续**。
- 单击“一般线性模型”对话框中的**估算的均值**。

图片 19-5
“一般线性模型：估算的均值”对话框



- 选择显示 shopfor、usecoup 和 shopfor*usecoup 交互的均值。
- 选择**简单**对比，并选择 **3 Self and family** 作为 shopfor 的参考类别。请注意，一旦选定，该类别在对话框中即显示为“3”。
- 选择**简单**对比，并选择 **1 No** 作为 usecoup 的参考类别。
- 单击**继续**。
- 单击“一般线性模型”对话框中的**确定**。

模型摘要

图片 19-6
R 方统计量

R 方	.601

a. 模型: Amount spent = (截取) + shopfor + usecoup + shopfor * usecoup

R 方，即判定系数，是模型拟合的程度度量。它表明模型说明了消费金额大约 60% 的变化，提供了很好的说明功能。您可能还希望向模型添加其他预测变量来进一步改进拟合。

模型效应检验

图片 19-7
主体间效应检验

源	df1	df2	Wald F	Sig.
(已校正的模型)	11.000	3.000	127.231	.001
(截取)	1.000	13.000	6321.597	.000
shopfor	2.000	12.000	643.593	.000
usecoup	3.000	11.000	87.453	.000
shopfor * usecoup	6.000	8.000	10.688	.002

a. 模型: Amount spent = (截取) + shopfor + usecoup + shopfor * usecoup

对模型的每一项以及模型整体进行检验，检验其效应的值是否等于 0。显著性值小于 0.05 的项具有一定可辨别效应。因此，所有模型项对模型都有贡献。

参数估计值

图片 19-8
参数估计值

参数	估计	标准误差	95% 置信区间		设计效果
			下限	下限	
(截取)	518.249	11.731	492.905	543.592	1.387
[shopfor=1]	-174.757	10.762	-198.006	-151.508	.950
[shopfor=2]	-129.443	11.455	-154.191	-104.696	.925
[shopfor=3]	.000[a]	.	.	.	.
[usecoup=1]	-140.838	10.180	-162.830	-118.846	.649
[usecoup=2]	-63.026	13.195	-91.531	-34.520	.940
[usecoup=3]	-31.375	9.726	-52.387	-10.363	.564
[usecoup=4]	.000[a]	.	.	.	.
[shopfor=1] * [usecoup=1]	41.693	11.170	17.562	65.824	.606
[shopfor=1] * [usecoup=2]	44.505	18.068	5.471	83.539	1.413
[shopfor=1] * [usecoup=3]	9.204	11.057	-14.684	33.092	.594
[shopfor=1] * [usecoup=4]	.000[a]	.	.	.	.
[shopfor=2] * [usecoup=1]	89.211	10.967	65.518	112.903	.533
[shopfor=2] * [usecoup=2]	54.267	14.949	21.972	86.562	.836
[shopfor=2] * [usecoup=3]	17.884	13.753	-11.828	47.595	.797
[shopfor=2] * [usecoup=4]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=1]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=2]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=3]	.000[a]	.	.	.	.
[shopfor=3] * [usecoup=4]	.000[a]	.	.	.	.

a. 设置为零，原因是此参数为冗余的。

b. 模型: Amount spent = (截取) + shopfor + usecoup + shopfor * usecoup

参数估计值显示每个预测变量对消费金额的效应。截距项的值 518.249 表明，对于使用报纸和目标邮件优惠券的家庭购物者，杂货连锁店可预期其平均消费 518.25 美元。可以看出，截距与这些因子水平关联，原因是这些因子水平的参数是冗余的。

- shopfor 系数表明，在同时使用邮件优惠券和报纸优惠券的顾客中，单身顾客的消费金额少于有配偶的顾客，后者的消费金额又少于需要抚养他人的顾客。由于模型效应检验显示此项对模型有贡献，这些差异就不是巧合。
- usecoup 系数表明，随着使用的优惠券数量的减少，需要抚养他人的顾客的消费金额会减少。估计值中存在一定的不确定性，但置信区间不包含 0。
- 交互系数表明，不使用优惠券或仅使用报纸优惠券的顾客，以及不需要抚养他人的顾客，消费金额大于预期值。如果某个交互参数有任何部分是冗余的，则该交互参数是冗余的。
- 设计效果的值与 1 之间的偏差表明，与假设这些观察值来自简单随机样本将获得的标准误相比，这些参数估计值的某些标准误计算值更大，另一些则更小。将抽样设计信息融入分析具有非常重要的意义，否则，举例来说，可能推断出 usecoup=3 系数与 0 没有区别！

参数估计值对于量化每个模型项都非常有用，但估计边际均值表可以更方便地解释模型结果。

估算边际均值

图片 19-9
基于“Who shopping for”水平的估计边际均值

Who shopping for	均值	标准误差	95% 置信区间	
			下限	下限
Self	308.5326	3.94286	300.0145	317.0506
Self and spouse	370.3361	4.87908	359.7955	380.8767
Self and family	459.4392	7.19769	443.8895	474.9888

该表显示 Amount spent 在 Who shopping for 因子水平上的模型估计边际均值和标准误。此表对于研究此因子的各个水平之间的差异很有用。 本示例中，单身顾客预期消费约 308.53 美元，有配偶的顾客预期消费 370.34 美元，需要抚养他人的顾客将消费 459.44 美元。要了解这代表真实差别还是巧合，请查看检验结果。

图片 19-10
对性别估计边际均值的单个检验结果

Who shopping for 的简单对比[a]	对比估计	假设值	差分（估计 - 已假设）	标准误差	df1	df2	Wald F	Sig.
水平 Self 与水平 Self and family	-150.907	.000	-150.907	4.903	1.000	13.000	947.409	.000
水平 Self and spouse 与水平 Self and family	-89.103	.000	-89.103	5.903	1.000	13.000	227.842	.000

a. 参考类别 = Self and family

单个检验表显示消费金额的两个简单对比。

- 对比估计值是所列 Who shopping for 水平消费金额的差异。
- 假设值 0.00 代表我们相信消费金额没有差异。
- 已显示自由度的 Wald F 统计量，用于检验对比估计值和假设值之间的差异是否归因于几率变异造成的。

- 显著性值小于 0.05，因此，可以得出结论，消费金额存在差异。

对比估计的值与参数估计值不相同。原因是，有一个交互项包含 Who shopping for 效应。因此，shopfor=1 的参数估计是 Self 和 Self and Family 水平在变量 Use coupons 的 From both 水平上的简单对比。此表中的对比估计在 Use coupons 的水平上取平均值。

图片 19-11
对性别估计边际均值的整体检验结果

df1	df2	Wald F	Sig.
2.000	12.000	643.593	.000

整体检验表报告单个检验表中所有对比的检验结果。其显著性值小于 0.05，证实在 Who shopping for 水平之间存在消费金额差异。

图片 19-12
基于购物风格水平的估计边际均值

Use coupons	均值	标准误差	95% 置信区间	
			下限	下限
No	319.6455	6.51429	305.5722	333.7188
From newspaper	386.7469	4.32295	377.4077	396.0861
From mailings	394.5028	5.54218	382.5297	406.4760
From both	416.8486	6.51260	402.7790	430.9182

该表显示 Amount spent 在 Use coupons 因子水平上的模型估计边际均值和标准误。此表对于研究此因子的各个水平之间的差异很有用。 本示例中，一个不使用优惠券的顾客预期消费金额约为 319.65 美元，一个使用优惠券的顾客预期消费金额则明显更多。

图片 19-13
对购物风格估计边际均值的单个检验结果

Use coupons 的简单对比[a]	对比估计	假设值	差分（估计 - 已假设）	标准误差	df1	df2	Wald F	Sig.
水平 From newspaper 与水平 No	67.101	.000	67.101	6.537	1.000	13.000	105.352	.000
水平 From mailings 与水平 No	74.857	.000	74.857	5.875	1.000	13.000	162.328	.000
水平 From both 与水平 No	97.203	.000	97.203	5.603	1.000	13.000	300.921	.000

a. 参考类别 = No

单个检验表显示三个简单对比，比较的是不使用优惠券的顾客和使用优惠券的顾客的消费金额。

检验的显著性值小于 0.05，因此，可以得出结论，使用优惠券的顾客比不使用优惠券的顾客的消费金额大。

图片 19-14
购物风格估计边际均值的整体检验结果

df1	df2	Wald F	Sig.
3.000	11.000	87.453	.000

整体检验表报告单个检验表中所有对比的检验结果。其显著性值小于 0.05，证实在 Use coupons 水平之间存在消费金额差异。请注意，由于假设对比值等于 0，Use coupons 和 Who shopping for 的整体检验等同于模型效应的检验。

图片 19-15
基于购物风格和性别水平的估计边际均值

Who shopping for	Use coupons	均值	标准误差	95% 置信区间	
				下限	下限
Self	No	244.3471	6.00949	231.3644	257.3298
	From newspaper	324.9708	5.94134	312.1353	337.8063
	From mailings	321.3207	4.11028	312.4410	330.2005
	From both	343.4916	6.57845	329.2797	357.7034
Self and spouse	No	337.1783	7.12181	321.7925	352.5640
	From newspaper	380.0468	7.91038	362.9574	397.1361
	From mailings	375.3141	6.22468	361.8665	388.7617
	From both	388.8054	7.12101	373.4214	404.1894
Self and family	No	377.4111	11.58215	352.3894	402.4328
	From newspaper	455.2232	6.14420	441.9494	468.4969
	From mailings	486.8736	10.76529	463.6166	510.1306
	From both	518.2488	11.73120	492.9050	543.5925

该表显示 Amount spent 在 Who shopping for 和 Use coupons 因子组合上的模型估计边际均值、标准误差和置信区间。此表对于研究这两个因子之间的交互效应很有用，这种效应存在于模型效应的检验中。

摘要

本示例中，估计边际均值表明 Who shopping for 和 Use coupons 不同水平上顾客消费金额之间的差异。模型效应的检验证实了这一点，也证实了 Who shopping for*Use coupons 交互效应的存在。模型摘要表表明存在的模型说明了数据中的大部分变差，通过添加更多预测变量，可以对模型进行改进。

相关过程

“复杂样本一般线性模型”过程是非常有用的工具，可以在根据复杂抽样设计抽取个案后对刻度变量进行建模。

- 复杂样本抽样向导用于指定复杂抽样设计指定项并获取一个样本。抽样向导创建的抽样计划文件包含一个缺省的分析计划，当您分析据此计划获取的样本时，可以在“计划”对话框中指定该抽样计划文件。
- 复杂样本分析准备向导用于为现有的复杂样本指定分析指定项。当您分析与该计划对应的样本时，可以在“计划”对话框中指定由抽样向导创建的分析计划文件。
- 使用复杂样本 Logistic 回归过程可以为分类响应建模。
- 使用复杂样本序数回归过程可以为序数响应建模。

章
20

复杂样本 Logistic 回归

“复杂样本 Logistic 回归”过程对通过复杂抽样方法抽取的样本的二元或多项因变量执行 logistic 回归分析。您还可以请求对子体进行分析。

使用复杂样本 Logistic 回归评估信用风险

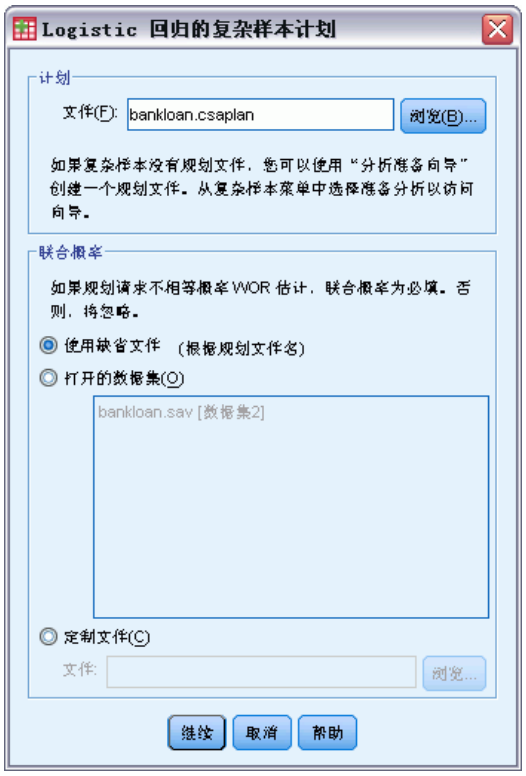
如果您是银行的信贷员，希望能够确定一些标志贷款者可能拖欠贷款的特征，然后使用这些特征确定良好信用风险和不良信用风险。

假设信贷员已根据一项复杂设计收集了客户过去在几个不同分支机构贷款的记录。该信息包含在 bankloan_cs.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件. 融入样本设计时，信贷员希望了解客户拖欠的概率是否与年龄、工作经历和信用负债量有关。

运行分析

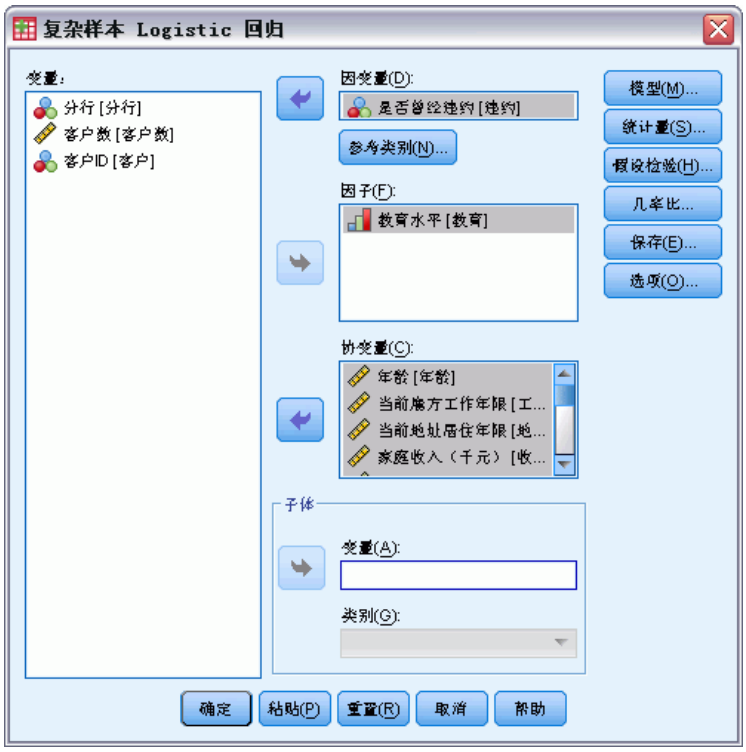
► 要创建 Logistic 回归模型，请从菜单中选择：
分析 > 复杂样本 > Logistic 回归...

图片 20-1
“复杂样本计划”对话框



- 浏览至 bankloan.csaplan 并将其选中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.
- 单击继续。

图片 20-2
“Logistic 回归”对话框



- 选择 Previously defaulted 作为因变量。
- 选择 Level of education 作为因子。
- 选择 Age in years 到 Other debt in thousands 作为协变量。
- 选择 Previously defaulted 并单击**参考类别**。

图片 20-3
“Logistic 回归: 参考类别”对话框

▶ 选择**最低值**作为参考类别。

此操作将“did not default”类别作为参考类别；因此，输出报告中几率比的属性为拖欠概率越大则几率比越大。

▶ 单击**继续**。

▶ 在“Logistic 回归”对话框中单击**统计量**。

图片 20-4
“Logistic 回归: 统计量”对话框

▶ 在“模型拟合度”组中选择**分类表**

▶ 选择“参数”组中的**估算**、**取幂估值**、**标准误**、**置信区间**以及**设计效应**。

▶ 单击**继续**。

▶ 在“Logistic 回归”对话框中单击**几率比**。

图片 20-5
“Logistic 回归：几率比”对话框

- 选择创建因子 ed 与协变量 employ 和 debtinc 的几率比。
- 单击**继续**。
- 在“Logistic 回归”对话框中单击**确定**。

伪 R 平方

图片 20-6
伪 R 平方统计量

Cox 和 Snell	.330
Nagelkerke	.451
McFadden	.304

因变量: Previously defaulted(参考类别 = No)
模型: (截取), ed, age, employ, address, income, debtinc, creddebt, othdebt

在线性回归模型中，判定系数 R^2 总结了因变量中与预测变量（自变量）关联的方差的比例，R^2 值越大表示模型解释的变异越多，最大为 1。对于使用分类因变量的回归模型，不可能计算在线性回归模型中具有所有 R^2 特性的单个 R^2 统计量，因此改为计算这些近似值。下面的方法用于估计判定系数。

- Cox & Snell R^2 (Cox 和 Snell, 1989) 基于该模型的对数似然估计，与基线模型的对数似然估计相对。但是，对于分类结果，其理论最大值小于 1，即使是对“完美”模型也是如此。
- Nagelkerke R^2 (Nagelkerke, 1991) 是 Cox & Snell R 方的调整版本，它调整了统计量的标度，以涵盖从 0 到 1 的完整范围。
- McFadden 的 R^2 (McFadden, 1974) 是另一个版本，它基于仅截距模型和完全估计的模型的对数似然估计内核。

构成“良好” R^2 值的内容在不同的应用领域之间各不相同。虽然这些统计量本身就有一定的意义，但是它们在与相同数据的竞争模型比较时最为有用。根据此度量，具有最大 R^2 统计量的模型为“最佳”。

Classification

图片 20-7
分类表

已观测	已预测		
	No	Yes	百分比更正
No	188289.67	31871.267	85.5%
Yes	49970.600	77675.133	60.9%
整体百分比	68.5%	31.5%	76.5%

因变量: Previously defaulted(参考类别 = No)
模型: (截取), ed, age, employ, address, income, debtinc, creddebt, othdebt

分类表显示使用 Logistic 回归模型的实际结果。 对于每个个案，只要个案的模型预测 logit 大于 0，预测响应就为 Yes。个案是由 finalweight 加权的，因此，分类表报告总体的期望模型性能。

- 对角线上的单元格是正确的预测值。
- 偏离对角线的单元格是不正确的预测值。

基于用于创建该模型的个案，可以期望对使用此模型的总体中 85.5% 的不拖欠者进行正确分类。同样，可以期望对 60.9% 的拖欠者进行正确分类。 在整体上，可以期望对 76.5% 的个案进行正确分类；但是，该表是由用于创建模型的个案构建的，因此，这些估计值可能过于乐观了。

模型效应检验

图片 20-8
主体间效应检验

源	df1	df2	Wald F	Sig.
(已校正的模型)	11.000	4.000	14.669	.010
(截取)	1.000	14.000	5.777	.031
ed	4.000	11.000	1.683	.224
age	1.000	14.000	5.352	.036
employ	1.000	14.000	88.244	.000
address	1.000	14.000	1.123	.307
income	1.000	14.000	.007	.932
debtinc	1.000	14.000	27.632	.000
creddebt	1.000	14.000	33.402	.000
othdebt	1.000	14.000	.709	.414

因变量: Previously defaulted(参考类别 = No)
模型: (截取), ed, age, employ, address, income, debtinc, creddebt, othdebt

对模型的每一项以及模型整体进行检验，检验其效应是否等于 0。显著性值小于 0.05 的项具有一定可辨别效应。因此，age、employ、debtinc 和 creddebt 对模型有贡献，而其他主效应则没有贡献。在数据的进一步分析中，考虑模型时，可以不包括 ed、address、income 和 othdebt。

参数估计值

图片 20-9
参数估计值

Previously defaulted	参数	B	标准误差	95% 置信区间 下限	95% 置信区间 下限	方根设计效果	Exp(B)	Exp(B) 的 95% 置信区间 下限	Exp(B) 的 95% 置信区间 下限
Yes	(截取)	-1.140	.399	-1.995	-.284	.815	.320	.136	.753
	[ed=1]	.720	.340	-.010	1.449	.929	2.054	.990	4.259
	[ed=2]	.684	.371	-.112	1.481	1.117	1.983	.894	4.397
	[ed=3]	.518	.307	-.140	1.177	.902	1.679	.869	3.244
	[ed=4]	.789	.302	.142	1.437	.904	2.202	1.152	4.208
	[ed=5]	.000[a]	.	.	.	.	1.000	.	.
	age	-.023	.010	-.043	-.002	.646	.978	.958	.998
	employ	-.225	.024	-.277	-.174	1.096	.798	.758	.840
	address	-.028	.026	-.085	.029	.807	.972	.919	1.029
	income	.000	.003	-.007	.006	1.187	1.000	.993	1.006
	debtinc	.095	.018	.056	.134	1.106	1.100	1.058	1.143
	creddebt	.493	.085	.310	.676	1.172	1.637	1.363	1.966
	othdebt	.026	.031	-.041	.094	1.104	1.027	.960	1.098

因变量: Previously defaulted(参考类别 = No)
模型: (截取), ed, age, employ, address, income, debtinc, creddebt, othdebt
a. 设置为零，原因是此参数为冗余的。

参数估计值表汇总了每个预测变量的作用。请注意：参数值影响“did default”类别相对于“did not default”类别的似然估计。因此，具有正系数的参数增大了拖欠的似然估计，而具有负系数的参数减小了拖欠的似然估计。

Logistic 回归系数的含义并不像线性回归系数的含义那样简单。尽管 B 便于检验模型效应，但 Exp(B) 却更容易解释。 对于不是交互项一部分的预测变量，Exp(B) 代表由于预测变量增加一个单位，事件几率的比率变化。例如，employ 的 Exp(B) 等于 0.798，意味着其他情况都相等的情况下，为其当前雇主工作两年的客户拖欠的几率是为其当前雇主工作一年的客户拖欠的几率的 0.798 倍。

设计效果表明，与假设这些观察值来自简单随机样本将获得的标准误相比，这些参数估计值的某些标准误计算值更大，另一些则更小。将抽样设计信息融入分析具有非常重要的意义，否则，举例来说，可能推断出年龄系数与 0 没有区别！

几率比

图片 20-10
教育水平的几率比

		Previously defaulted	几率比	95% 置信区间 下限	95% 置信区间 下限
Level of education	Did not complete high school 与 Post-undergraduate degree	Yes	2.054	.990	4.259
	High school degree 与	Yes	1.983	.894	4.397
	Some college 与	Yes	1.679	.869	3.244
	College degree 与	Yes	2.202	1.152	4.208

因变量: Previously defaulted(参考类别 = No)
模型: (截取), ed, age, employ, address, income, debtinc, creddebt, othdebt
a. 计算中使用的因子和协变量固定为以下值: Level of education=Post-undergraduate degree; Age in years=34.19; Years with current employer=6.99; Years at current address=6.32; Household income in thousands=60.1581; Debt to income ratio (x100)=9.9341; Credit card debt in thousands=1.9764; Other debt in thousands=3.9164

此表显示 Previously defaulted 在因子水平 Level of education 上的几率比。报告值是 Did not complete high school 到 College degree 拖欠几率与 Post-undergraduate degree 拖欠几率之比。因此，表中第一行中的几率比 2.054 表示未完成高中学业的客户的拖欠几率是具有研究生学历的客户的拖欠几率的 2.054 倍。

图片 20-11
为当前雇主工作年限的几率比

更改单位		Previously defaulted	几率比	95% 置信区间	
				下限	下限
Years with current employer	1.000	Yes	.798	.758	.840

因变量: Previously defaulted(参考类别 = No)
模型: (截取), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. 计算中使用的因子和协变量固定为以下值: Level of education=Post-undergraduate degree; Age in years=34.19; Years with current employer=6.99; Years at current address=6.32; Household income in thousands=60.1581; Debt to income ratio (x100)=9.9341; Credit card debt in thousands=1.9764; Other debt in thousands=3.9164

此表显示协变量 Years with current employer 变化一个单位时 Previously defaulted 的几率比。报告值是为当前雇主工作 7.99 年的客户拖欠几率与工作 6.99 年的客户拖欠几率之比（均值）。

图片 20-12
负债收入比的几率比

更改单位		Previously defaulted	几率比	95% 置信区间	
				下限	下限
Debt to income ratio (x100)	1.000	Yes	1.100	1.058	1.143

因变量: Previously defaulted(参考类别 = No)
模型: (截取), ed, age, employ, address, income, debtinc, creddebt, othdebt

a. 计算中使用的因子和协变量固定为以下值: Level of education=Post-undergraduate degree; Age in years=34.19; Years with current employer=6.99; Years at current address=6.32; Household income in thousands=60.1581; Debt to income ratio (x100)=9.9341; Credit card debt in thousands=1.9764; Other debt in thousands=3.9164

此表显示协变量 Debt to income ratio 变化一个单位时 Previously defaulted 的几率比。报告值是负债/收入比为 10.9341 的客户拖欠几率与负债/收入比为 9.9341 的客户拖欠几率之比（均值）。

请注意：这些预测变量都不是交互项的部分，因此，这些表中报告的几率比的值等于取幂参数估计值。如果某个预测变量是交互项的部分，则这些表中报告的该变量的几率比还将取决于构成交互的其他预测变量的值。

摘要

已使用复杂样本 Logistic 回归过程构建了一个模型，将用于预测给定客户可能拖欠贷款的概率。

信贷员所面临的严重问题是 I 类和 II 类错误带来的损失。也就是说，将拖欠贷款者归类为未拖欠贷款者会造成什么损失（I 类）？将未拖欠贷款者归类为拖欠贷款者会造成什么损失（II 类）？如果坏帐是主要问题，您会希望减少 I 类错误并尽量提高**敏感度**。如果首要任务是扩展客户群，您会希望减少 II 类错误并尽量提高**特异性**。通常情况下，两者都是重要问题，因此需要选择客户分类的决策规则，使敏感度和特异性达到最佳配合。

相关过程

“复杂样本 Logistic 回归”过程是非常有用的工具，可以在根据复杂抽样设计抽取个案后对分类变量进行建模。

- 复杂样本抽样向导用于指定复杂抽样设计指定项并获取一个样本。抽样向导创建的抽样计划文件包含一个缺省的分析计划，当您分析据此计划获取的样本时，可以在“计划”对话框中指定该抽样计划文件。
- 复杂样本分析准备向导用于为现有的复杂样本指定分析指定项。当您分析与该计划对应的样本时，可以在“计划”对话框中指定由抽样向导创建的分析计划文件。
- 使用复杂样本一般线性模型过程可以为刻度响应建模。
- 使用复杂样本序数回归过程可以为序数响应建模。

章
21

复杂样本序数回归

“复杂样本序数回归”过程为通过复杂抽样方法抽取的样本的序数因变量创建预测模型。您还可以请求对子体进行分析。

使用复杂样本序数回归分析调查结果

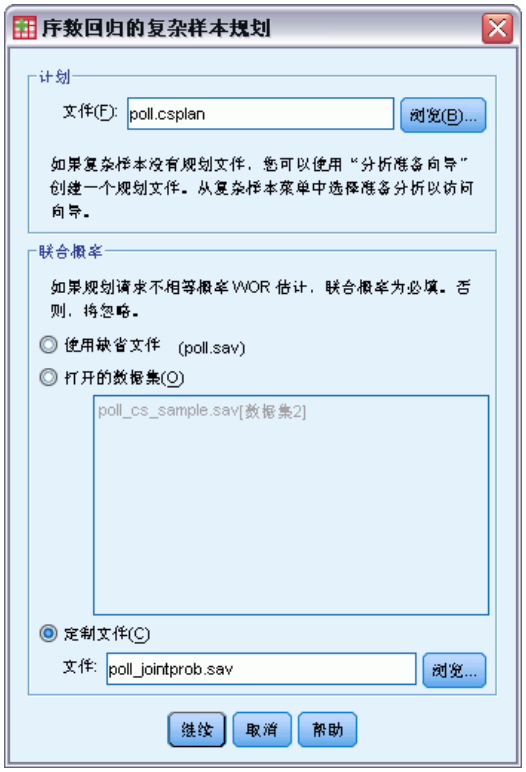
议员在向立法院提交某项法案之前想了解公众是否支持该法案，以及对该法案的支持与选民人群统计信息有何关联。民意测验专家根据复杂抽样设计并实施了一些采访。

调查结果收集在 poll_cs_sample.sav中。民意测验专家使用的抽样计划包含在 poll.csplan 中；该计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (poll_jointprob.sav) 包含联合选择概率。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.使用复杂样本序数回归，根据选民人群统计信息，将模型与对法案的支持水平进行拟合。

运行分析

▶ 要运行复杂样本序数回归分析，请从菜单中选择：
分析 > 复杂样本 > 序数回归...

图片 21-1
“复杂样本计划”对话框



- 浏览至 poll.csplan 并选择其作为计划文件。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.
- 选择 poll_jointprob.sav 作为联合概率文件。
- 单击继续。

图片 21-2
“Ordinal 回归”对话框

- 选择立法机构应该颁布汽油税作为因变量。
- 选择年龄分段到驾驶频率作为因子。
- 单击**统计量**。

图片 21-3
“Ordinal 回归：统计量”对话框

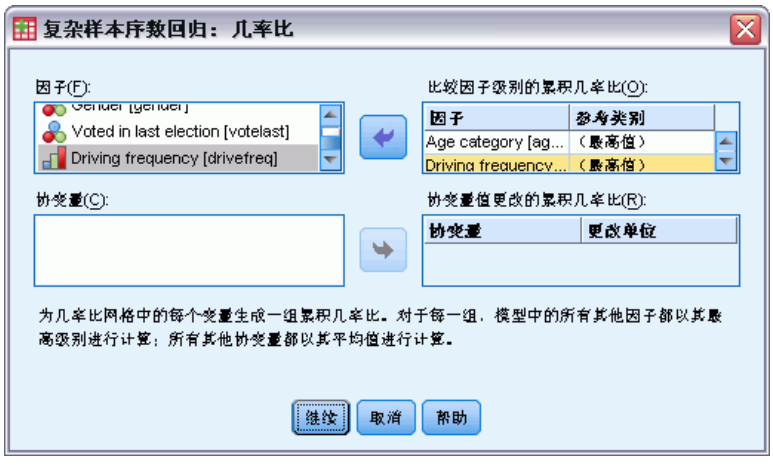
- 在“模型拟合度”组中选择**分类表**。
- 选择“参数”组中的**估算**、**取幂估值**、**标准误**、**置信区间**以及**设计效应**。
- 选择**相等斜率的 Wald 检验**和**广义(不等斜率)模型的参数估计**。
- 单击**继续**。
- 在“复杂样本序数回归”对话框中单击**假设检验**。

图片 21-4
“假设检验”对话框

对于平行线检验而言，即使是适量的预测变量和响应类别，Wald F 检验统计量也可能是无法估计的。

- 选择“检验统计”组中的**调整 F**。
- 选择**连续 Sidak** 作为多比较的调整方法。
- 单击**继续**。
- 在“复杂样本序数回归”对话框中单击**几率比**。

图片 21-5
“Ordinal 回归：几率比”对话框



- 选择生成年龄分段和驾驶频率的累积几率比。

- ▶ 选择 10-14,999 里/年，一个比最大值更为典型的年里程，作为驾驶频率的参考类别。
- ▶ 单击**继续**。
- ▶ 在“复杂样本序数回归”对话框中单击**确定**。

伪 R 平方

图片 21-6
伪 R 平方

Cox 和 Snell	.179
Nagelkerke	.191
McFadden	.071

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, drivefreq
关联函数: Logit

在线性回归模型中，判定系数 R^2 总结了因变量中与预测变量（自变量）关联的方差的比例，R^2 值越大表示模型解释的变异越多，最大为 1。对于使用分类因变量的回归模型，不可能计算在线性回归模型中具有所有 R^2 特性的单个 R^2 统计量，因此改为计算这些近似值。下面的方法用于估计判定系数。

- Cox & Snell R^2(Cox 和 Snell, 1989)基于该模型的对数似然估计，与基线模型的对数似然估计相对。但是，对于分类结果，其理论最大值小于 1，即使是对“完美”模型也是如此。
- Nagelkerke R^2(Nagelkerke, 1991)是 Cox & Snell R 方的调整版本，它调整了统计量的标度，以涵盖从 0 到 1 的完整范围。
- McFadden 的 R^2(McFadden, 1974)是另一个版本，它基于仅截距模型和完全估计的模型的对数似然估计内核。

构成 “良好” R^2 值的内容在不同的应用领域之间各不相同。虽然这些统计量本身就有一定的意义，但是它们在与相同数据的竞争模型比较时最为有用。根据此度量，具有最大 R^2 统计量的模型为“最佳”。

模型效应检验

图片 21-7
模型效应检验

源	df1	df2	调整的 Wald F	Sig.	序列 Sidak Sig.
agecat	2.283	31.966	6.215	.004	.003
gender	1.000	14.000	.046	.834	.834
votelast	1.000	14.000	.076	.787	.787
drivefreq	3.785	52.987	228.015	.000	.000

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, gender, votelast, drivefreq
关联函数: Logit

检验模型每一项的效应是否等于 0。显著性值小于 0.05 的项具有定可辨别效应。因此，agecat 和 drivefreq 对模型有贡献，而其他主效应则没有。进一步分析数据时，可以将 gender 和 votelast 排除在模型之外。

参数估计值

参数估计值表汇总了每个预测变量的作用。关联函数的性质决定了此模型的系数难于解释，但是，通过协变量系数的符号和因子水平系数的相对值，还是可以获得模型预测变量效应的重要信息。

- 对于协变量，正（负）系数表示预测变量和结果的正（反）关系。系数为正的协变量值增加，则对应于处于“更高”累积结果类别的概率增加。
- 对于因子，系数越大的因子水平表示处于“更高”累积结果类别的概率越大。因子水平的系数符号取决于因子水平相对于参考类别的效应。

图片 21-8
参数估计值

参数		B	标准误差	95% 置信区间		方根设计效果	Exp(B)	Exp(B) 的 95% 置信区间	
				下限	下限			下限	下限
阈值	[opinion_gastax=1]	-3.343	.104	-3.566	-3.120	1.064	.035	.028	.044
	[opinion_gastax=2]	-1.910	.098	-2.120	-1.700	1.029	.148	.120	.183
	[opinion_gastax=3]	-.674	.090	-.866	-.482	.957	.510	.421	.618
回归	[agecat=1]	-.324	.079	-.494	-.154	1.339	.723	.610	.858
	[agecat=2]	-.138	.054	-.255	-.022	1.076	.871	.775	.978
	[agecat=3]	-.095	.076	-.257	.068	1.485	.909	.773	1.070
	[agecat=4]	.000[a]	.	.	.	.	1.000	.	.
	[gender=0]	-.008	.035	-.084	.068	.974	.992	.920	1.071
	[gender=1]	.000[a]	.	.	.	.	1.000	.	.
	[votelast=0]	-.011	.039	-.095	.073	1.050	.989	.909	1.076
	[votelast=1]	.000[a]	.	.	.	.	1.000	.	.
	[drivefreq=1]	-3.751	.153	-4.079	-3.423	1.057	.023	.017	.033
	[drivefreq=2]	-3.003	.116	-3.251	-2.755	1.107	.050	.039	.064
	[drivefreq=3]	-2.295	.114	-2.540	-2.050	1.259	.101	.079	.129
	[drivefreq=4]	-1.570	.092	-1.769	-1.372	1.038	.208	.171	.254
	[drivefreq=5]	-.812	.089	-1.003	-.621	.970	.444	.367	.537
	[drivefreq=6]	.000[a]	.	.	.	.	1.000	.	.

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, gender, votelast, drivefreq
关联函数: Logit

a. 设置为零，原因是此参数为冗余的。

根据参数估计值，可以进行以下解释：

- 年龄组较低的人比最高年龄组的人更支持该法案。
- 较少开车的人比经常开车的人更支持法案。
- 变量 gender 和 votelast 的系数除了统计显著性低之外，还小于其他系数。

设计效果表明，与使用简单随机样本将获得的标准误相比，这些参数估计值的部分标准误计算值更大，另一些则更小。将抽样设计信息融入分析具有非常重要的意义，否则，举例来说，可能推断出年龄组的第三水平 [agecat=3] 的系数与 0 显著不同！

Classification

图片 21-9
分类变量信息

		加权计数	加权百分比
The legislature should enact a gas tax	Strongly agree	25132.955	21.3%
	Agree	32261.425	27.3%
	Disagree	29477.417	24.9%
	Strongly disagree	31314.203	26.5%
Age category	18-30	20509.504	17.4%
	31-45	35380.506	29.9%
	46-60	34865.792	29.5%
	>60	27430.198	23.2%
Gender	Male	61424.547	52.0%
	Female	56761.453	48.0%
Voted in last election	No	70607.216	59.7%
	Yes	47578.784	40.3%
Driving frequency	Do not own car	3437.137	2.9%
	<10,000 miles/year	10816.349	9.2%
	10-14,999 miles/year	32539.364	27.5%
	15-19,999 miles/year	39179.814	33.2%
	20-29,999 miles/year	25617.804	21.7%
	>=30,000 miles/year	6595.532	5.6%
种群大小		118186.00	100.0%

a. 因变量值按升序排序。

给定观察数据的“空”模型（即没有预测变量的模型）会将所有客户分类为模态组支持。因此，空模型将修正 27.3% 的时间。

图片 21-10
分类表

	已预测				
已观测	Strongly agree	Agree	Disagree	Strongly disagree	百分比更正
Strongly agree	7067.567	12130.814	3875.825	2058.750	28.1%
Agree	4271.234	14464.286	7320.767	6205.137	44.8%
Disagree	2024.816	11703.368	7108.487	8640.746	24.1%
Strongly disagree	889.869	8169.109	6946.522	15308.703	48.9%
整体百分比	12.1%	39.3%	21.4%	27.3%	37.2%

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, gender, votelast, drivefreq
关联函数: Logit

分类表显示使用模型的实际结果。对于每个个案，预测响应都是模型预测概率最高的响应类别。个案由 Final Sampling Weight 加权，以便分类表报告总体的期望模型性能。

- 对角线上的单元格是正确的预测值。
- 偏离对角线的单元格是不正确的预测值。

模型的正确分类提高了 9.9%，即 37.2% 的个案。具体来说，模型对 Agree 或 Strongly disagree 的选民的分类性能有相当的提高，对 Disagree 的选民的分类性能略有下降。

几率比

累积几率定义为因变量取值小于或等于给定响应类别的概率与大于该响应类别的概率的比率。**累积几率比**是不同预测变量值的累积几率之比，与取幂参数估计值紧密相关。有趣的是，累积几率比本身不取决于响应类别。

图片 21-11
年龄组的累积几率比

		累积几率比	95% 置信区间		设计效果	方根设计效果
			下限	下限		
Age category	18-30 与 >60	1.383	1.166	1.639	1.793	1.339
	31-45 与 >60	1.148	1.022	1.290	1.158	1.076
	46-60 与 >60	1.100	.935	1.294	2.206	1.485

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, gender, votelast, drivefreq
关联函数: Logit

a. 计算中使用的因子和协变量固定为以下值: Age category=>60; Gender=Female; Voted in last election=Yes; Driving frequency=>=30,000 miles/year

此表显示 Age category 的因子水平的累积几率比。报告值是 18 - 30 到 46 - 60 与 >60 的累积几率的比率。因此，表中第一行的几率比 1.383 意味着年龄为 18 - 30 的选民的累积几率是大于 60 岁的选民的 1.383 倍。请注意，Age category 与任何交互项无关，因此，这些几率比只是取幂参数估计值的比率。例如，18 - 30 对 >60 的累积几率比为 1.00 / 0.723 = 1.383。

图片 21-12
开车频率的几率比

		累积几率比	95% 置信区间		设计效果	方根设计效果
			下限	下限		
Driving frequency	Do not own car 与 10-14,999 miles/year	4.288	2.878	6.390	2.345	1.531
	<10,000 miles/year 与 10-14,999 miles/year	2.030	1.656	2.488	1.838	1.356
	15-19,999 miles/year 与 10-14,999 miles/year	.484	.430	.546	1.450	1.204
	20-29,999 miles/year 与 10-14,999 miles/year	.227	.193	.267	2.095	1.448
	>=30,000 miles/year 与 10-14,999 miles/year	.101	.079	.129	1.585	1.259

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, gender, votelast, drivefreq
关联函数: Logit

a. 计算中使用的因子和协变量固定为以下值: Age category=>60; Gender=Female; Voted in last election=Yes; Driving frequency=>=30,000 miles/year

此表显示 Driving frequency 的因子水平的累积几率比，使用 10 - 14,999 miles/year 作为参考类别。Driving frequency 与任何交互项无关，因此，这些几率比只是取幂参数估计值的比率。例如，20 - 29,999 miles/year对10 - 14,999 miles/year的累积几率比为 0.101/0.444 = 0.227。

一般化累积模型

图片 21-13
平行线检验

df1	df2	调整的 Wald F	Sig.	序列 Sidak Sig.
8.769	122.767	1.894	.061	.392

因变量:The legislature should enact a gas tax (引序)
模型:(阈值), agecat, gender, votelast, drivefreq
关联函数: Logit

平行线检验可帮助评估所有响应类别的参数都相同这一假设是否合理。此检验将所有类别都具有一组系数的估计模型与每个类别都具有一组单独系数的一般化模型进行比较。

Wald F 检验是对平行线假设的对比矩阵进行的 Omnibus 检验，该假设提供渐近修正的 p 值；对于小到中等大小的样本，调整的 Wald F 统计量性能很好。显著性值接近 0.05，说明一般化模型可改善模型拟合；但是，顺序 Sidak 调整检验报告的显著性值 (0.392) 足够高，整体上，没有明确的证据可拒绝平行线假设。顺序 Sidak 检验从单个对比 Wald 检验开始，提供一个整体 p 值，这些结果应与 Omnibus Wald 检验结果相当。本示例中，它们的差异之大令人吃惊，但这是因为检验中存在很多对比，并且设计自由度相对较小。

图片 21-14
一般化累积模型的参数估计值（显示部分）

The legislature should enact a gas tax	参数	B	标准误差	95% 置信区间 下限	下限
Strongly agree	(阈值)	-3.681	.221	-4.155	-3.207
	[agecat=1]	-.320	.096	-.525	-.115
	[agecat=2]	-.075	.071	-.227	.077
	[agecat=3]	-.022	.073	-.180	.135
	[agecat=4]	.000[a]	.		
	[gender=0]	-.082	.054	-.197	.033
	[gender=1]	.000[a]	.		
	[votelast=0]	.008	.052	-.104	.120
	[votelast=1]	.000[a]	.		
	[drivefreq=1]	-4.096	.267	-4.669	-3.523
	[drivefreq=2]	-3.367	.237	-3.876	-2.857
	[drivefreq=3]	-2.678	.224	-3.158	-2.199
	[drivefreq=4]	-1.928	.213	-2.384	-1.471
	[drivefreq=5]	-1.015	.252	-1.555	-.476
	[drivefreq=6]	.000[a]	.		
Agree	(阈值)	-1.963	.153	-2.291	-1.635
	[agecat=1]	-.385	.095	-.587	-.182
	[agecat=2]	-.130	.069	-.279	.018
	[agecat=3]	-.139	.101	-.356	.077
	[agecat=4]	.000[a]	.		
	[gender=0]	-.004	.040	-.090	.082
	[gender=1]	.000[a]	.		
	[votelast=0]	.009	.059	-.117	.135
	[votelast=1]	.000[a]	.		
	[drivefreq=1]	-3.867	.318	-4.549	-3.185
	[drivefreq=2]	-3.005	.175	-3.380	-2.630
	[drivefreq=3]	-2.290	.187	-2.691	-1.888
	[drivefreq=4]	-1.633	.166	-1.988	-1.278
	[drivefreq=5]	-.909	.137	-1.204	-.615
	[drivefreq=6]	.000[a]	.		

此外，一般化模型系数的估计值与平行线假设下的估计值相差不大。

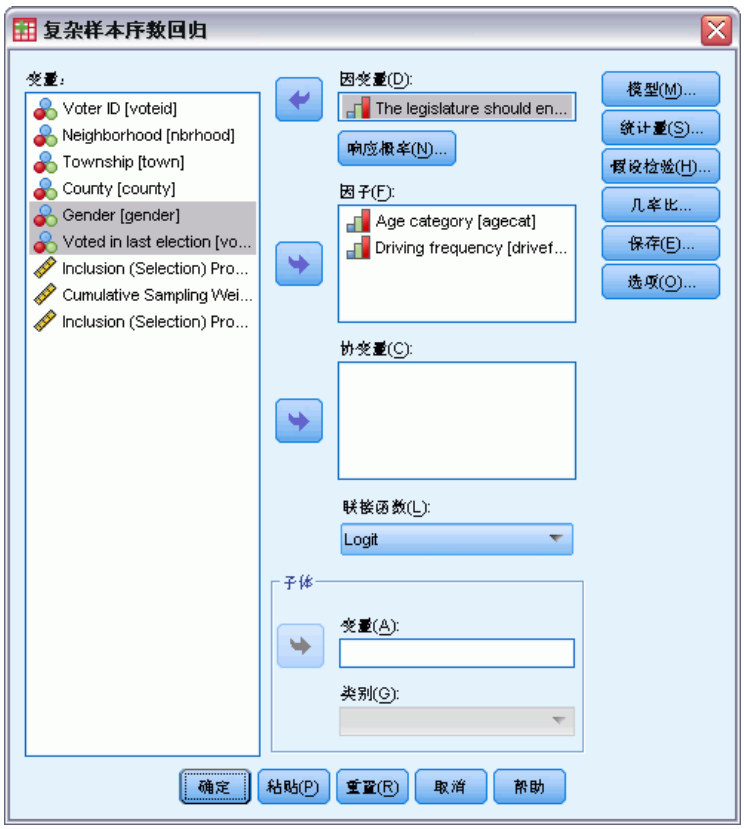
减少非显著性预测变量

模型效应的检验显示，性别和最终选择结果的模型系数与 0 没有显著的统计差别。

▶ 要生成简化模型，请调用“复杂样本序数回归”对话框。

▶ 单击“计划”对话框中的继续。

图片 21-15
“Ordinal 回归”对话框



▶ 取消选择性别和最终选择结果作为因子。

▶ 单击选项。

图片 21-16
Ordinal 回归：选项对话框

- 选择显示迭代历史记录。

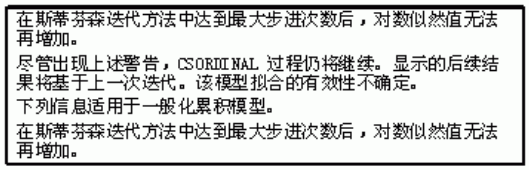
  迭代历史记录对于诊断估计算法遇到的问题很有用。

- 单击继续。

- 在“复杂样本序数回归”对话框中单击确定。

警告

图片 21-17
简化模型的警告



警告提示，简化模型的估计在参数估计值收敛之前结束，原因是对数似然估计不随当前参数估计值的任何变化（即“步进”）而增加。

图片 21-18
简化模型的警告

迭代数目[b]	N 斯蒂芬森迭代	伪 -2 对数似然	阈值			回归							
			[opinion_gastax=1]	[opinion_gastax=2]	[opinion_gastax=3]	[agecat=1]	[agecat=2]	[agecat=3]	[drivefreq=1]	[drivefreq=2]	[drivefreq=3]	[drivefreq=4]	[drivefreq=5]
0	0	326640.3	-1.309	-.058	1.020	.000	.000	.000	.000	.000	.000	.000	.000
1	0	303567.5	-3.242	-1.881	-.704	-.323	-.137	-.094	-3.841	-2.970	-2.248	-1.563	-.835
2	0	303336.3	-3.327	-1.897	-.664	-.325	-.139	-.095	-3.740	-2.998	-2.291	-1.568	-.811
3	0	303335.9	-3.333	-1.900	-.664	-.326	-.139	-.096	-3.750	-3.003	-2.295	-1.570	-.812
4	0	303335.9	-3.333	-1.900	-.664	-.326	-.139	-.096	-3.750	-3.003	-2.295	-1.570	-.812
5[a]	5	303335.9	-3.333	-1.900	-.664	-.326	-.139	-.096	-3.750	-3.003	-2.295	-1.570	-.812

不显示冗余参数。这些参数的值在所有迭代中总是为零。
因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, drivefreq
关联函数: Logit

a. 在斯蒂芬森迭代方法中达到最大步进次数后，对数似然值无法再增加。

b. 使用了 Newton-Raphson 方法来估计参数。

查看迭代历史记录，参数估计值在前几次迭代中的变化很小，可以不用特别注意该警告消息。

比较模型

图片 21-19
简化模型的伪 R 平方

Cox 和 Snell	.179
Nagelkerke	.191
McFadden	.071

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, drivefreq
关联函数: Logit

简化模型与原始模型的 R^2 值相同。这证明了简化模型的优点。

图片 21-20
简化模型的分类表

已观测	已预测				
	Strongly agree	Agree	Disagree	Strongly disagree	百分比更正
Strongly agree	7067.567	12823.258	3183.380	2058.750	28.1%
Agree	4271.234	15684.090	6100.963	6205.137	48.6%
Disagree	2024.816	13157.809	5654.047	8640.746	19.2%
Strongly disagree	889.869	9226.578	5889.053	15308.703	48.9%
整体百分比	12.1%	43.1%	17.6%	27.3%	37.0%

因变量:The legislature should enact a gas tax (升序)
模型:(阈值), agecat, drivefreq
关联函数: Logit

分类表在一定程度上使问题复杂化。简化模型的整体分类率 37.0% 与原始模型相当，这证明了简化模型的优点。但是，简化模型使 3.8% 的选民的预测响应从 Disagree 转移到了 Agree，这部分选民中，多数响应的是 Disagree 或 Strongly disagree。这个区别非常重要，需要在选择简化模型之前认真考虑。

摘要

已使用“复杂样本序数回归”过程，根据选民人群统计信息为提交的法案的支持水平构建了竞争模型。平行线检验显示一般化累积模型不是必要的。模型效应的检验表明，可从模型中减少 Gender 和 Voted in last election，与原始模型相比，简化模型的伪 R^2 和整体分类率性能良好。但是，简化模型在划分 Agree/Disagree 时错分的选民会增加，因此，现在立法者倾向保留原始模型。

相关过程

“复杂样本序数回归”过程是有用的工具，可以在根据复杂抽样设计抽取个案后对序数变量进行建模。

- 复杂样本抽样向导用于指定复杂抽样设计指定项并获取一个样本。抽样向导创建的抽样计划文件包含一个缺省的分析计划，当您分析据此计划获取的样本时，可以在“计划”对话框中指定该抽样计划文件。
- 复杂样本分析准备向导用于为现有的复杂样本指定分析指定项。当您分析与该计划对应的样本时，可以在“计划”对话框中指定由抽样向导创建的分析计划文件。
- 使用复杂样本一般线性模型过程可以为刻度响应建模。
- 使用复杂样本 Logistic 回归过程可以为分类响应建模。

章
22

复杂样本 Cox 回归

复杂样本 Cox 回归过程对由复杂取样方法抽取的样本进行生存分析。

在复杂样本 Cox 回归中使用依时预测器

政府执法机构关心其管辖区域内的屡犯率。测量屡犯率的方法之一就是罪犯第二次被捕的时间。机构希望利用按照复杂抽样方式绘制的样本上的 Cox 回归对时间建模以进行再次抓捕，但又担心成比例的风险假定在跨越年龄类别时失效。

从抽样部门选取 2003 年 6 月释放的第一次犯罪的罪犯，及至 2006 年 6 月底对其进行的个案历史调查。样本收集在 recidivism_cs_sample.sav 中。使用的抽样计划包含在 recidivism_cs.csplan 中；该计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (recidivism_cs_jointprob.sav) 包含联合选择概率。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.使用复杂样本 Cox 回归评估成比例的风险假设的有效性，且如果合适的话，使用依时预测器拟合模型。

准备数据

数据集包含第一次逮捕和第二次逮捕的释放日期；由于 Cox 回归分析了存活时间，您需要计算这些日期之间的时间量。

但是，第二次逮捕的日期 [date2] 包含具有 10/03/1582 值（日期变量缺失值）的个案。这些个案为没有第二次犯罪的人，我们当然想将其包括在模型中的右侧已审查个案中。追踪时期的结束日为 2006 年 6 月 30 日，所以我们将把 10/03/1582 重新编码为 06/30/2006。

▶ 要对这些值进行重新编码，请从菜单中选择：
转换 > 计算变量...

图片 22-1
“计算变量”对话框

- 键入 date2 作为目标变量。
- 键入 DATE.DMY(30,6,2006) 作为表达式。
- 单击如果。

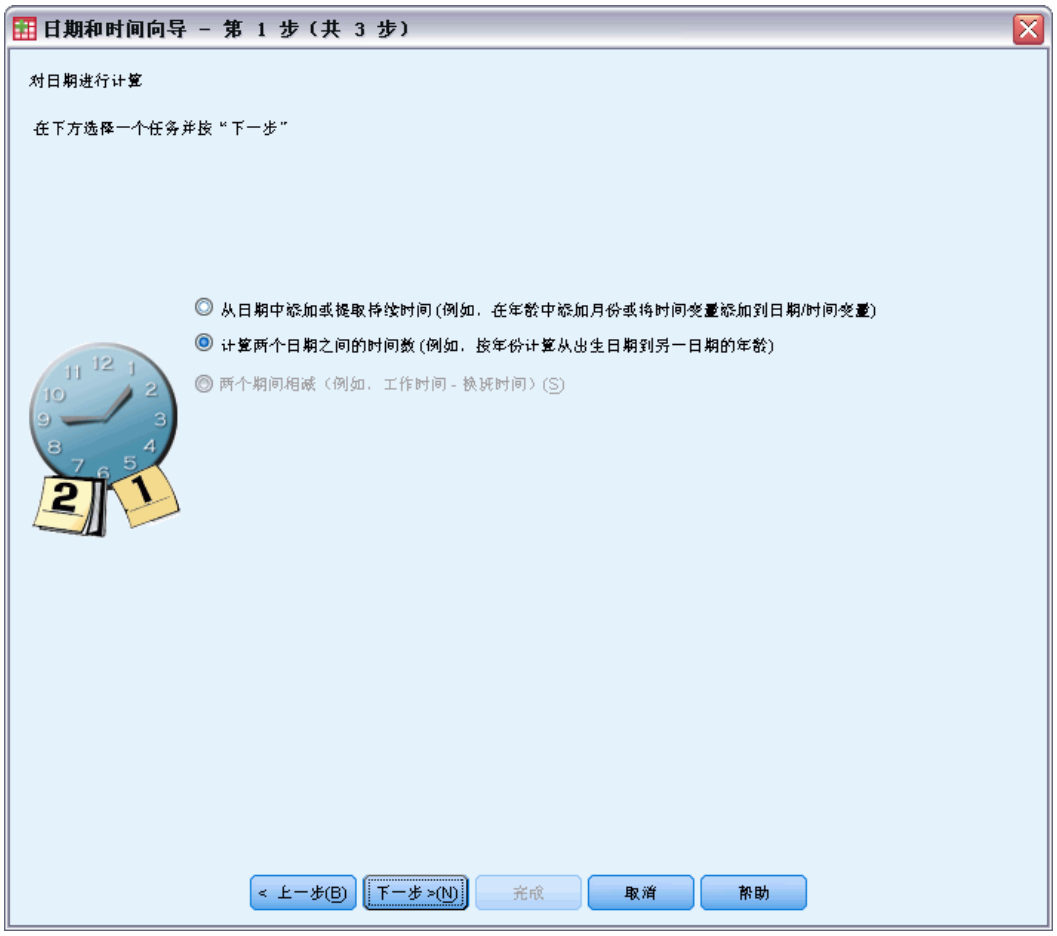
图片 22-2
“计算变量：If 个案”对话框

- 选中**如果个案满足条件则包含**。
- 键入 MISSING(date2) 作为表达式。
- 单击**继续**。
- 在“计算变量”对话框中单击**确定**。
- 下一步，要计算第一次逮捕和第二次逮捕之间的时间，请从菜单中选择：
  转换 > 日期和时间向导...

图片 22-3
日期和时间向导，“欢迎”步骤

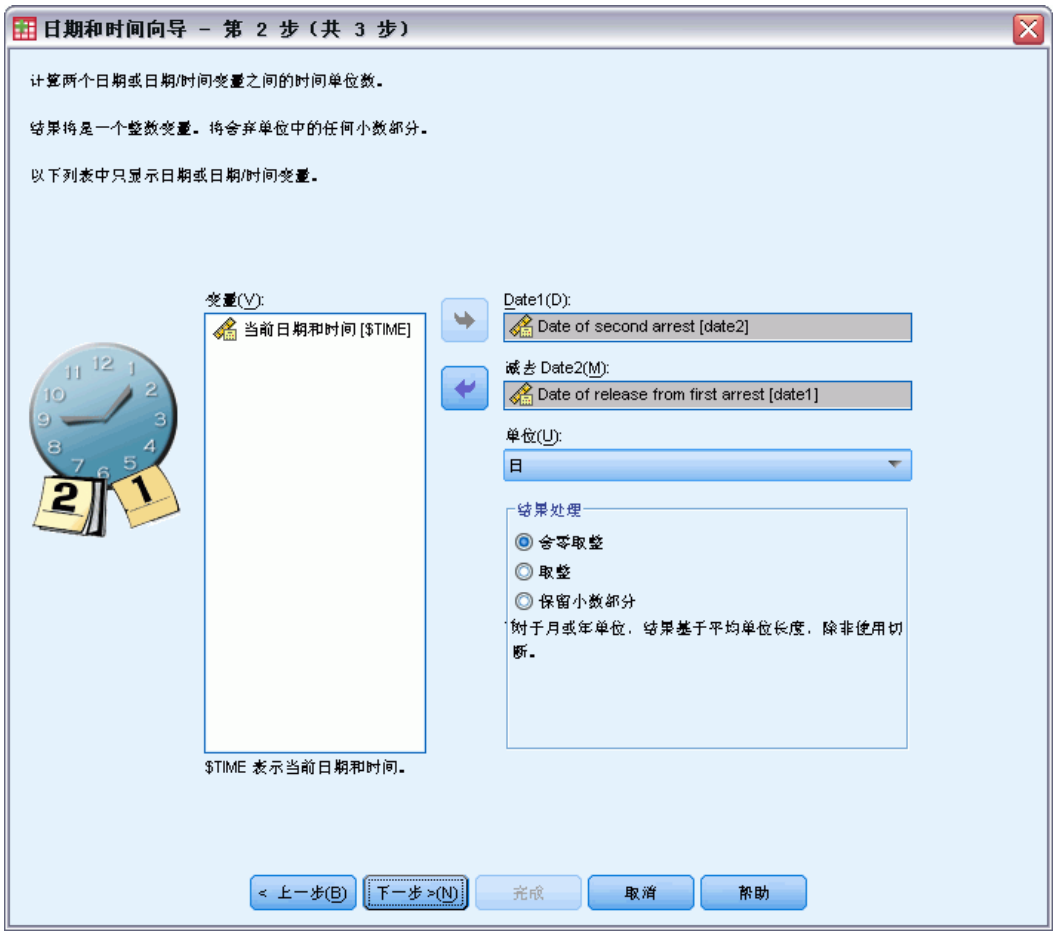
- 选择使用日期和时间进行计算。
- 单击下一步。

图片 22-4
日期和时间向导，“计算日期”步骤



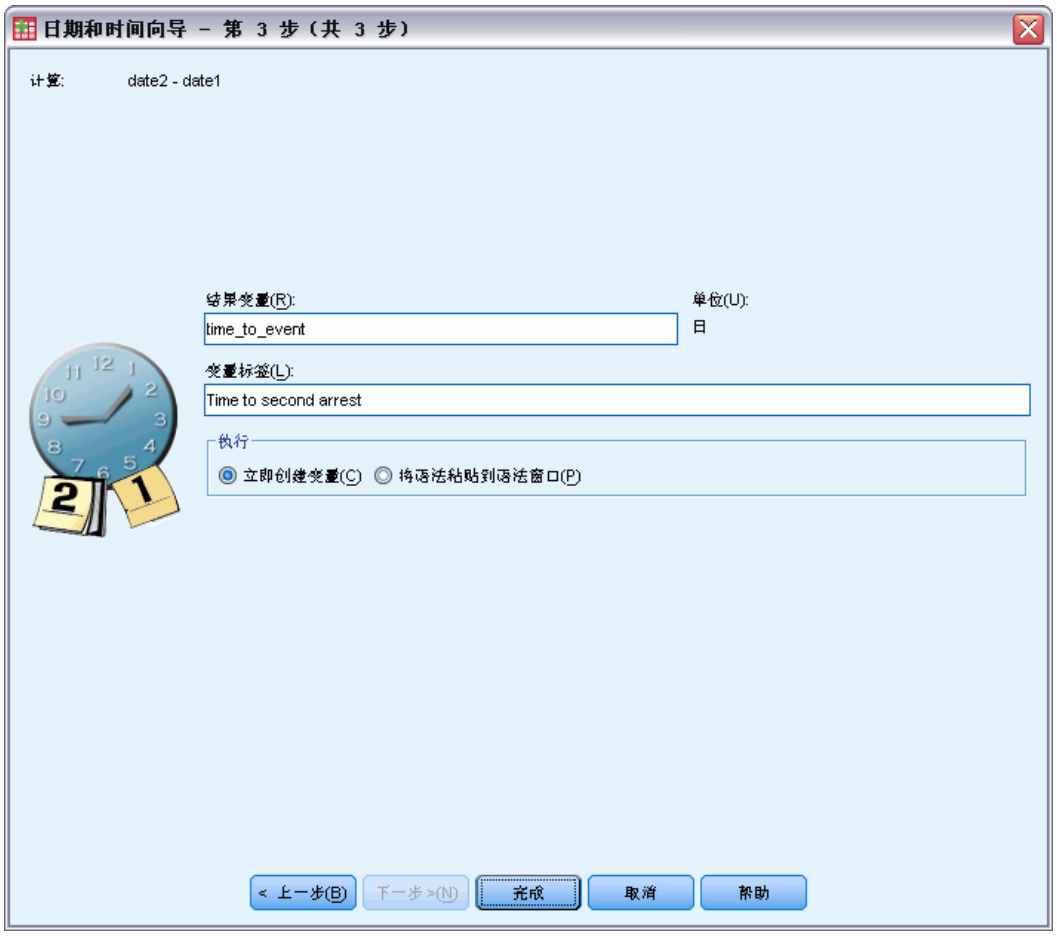
- 选择计算两个日期之间的时间单位数。
- 单击下一步。

图片 22-5
日期和时间向导，“计算两个日期之间的时间单位数”步骤



- 选择 第二次逮捕日期 [date2] 作为第一日期。
- 选择 第一次逮捕释放日期 [date1] 作为要从第一日期减去的日期。
- 选择**天数**作为单位。
- 单击**下一步**。

图片 22-6
日期和时间向导，“计算”步骤

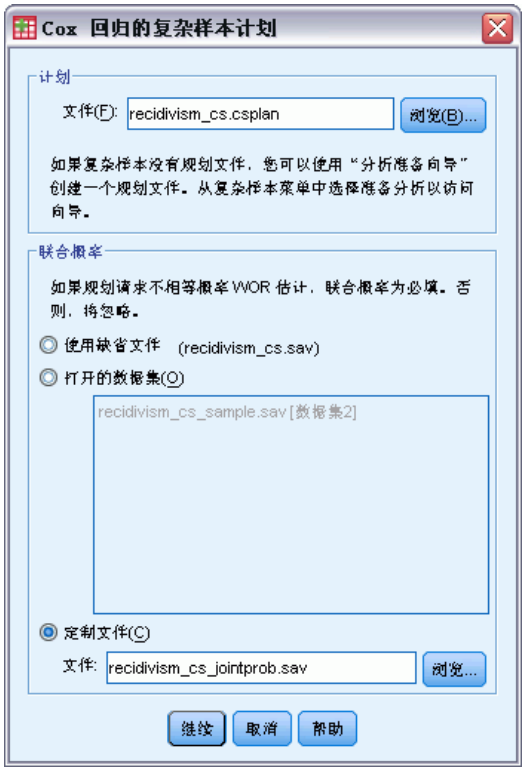


- 输入 time_to_event 作为代表两个日期之间的时间的变量的名称。
- 输入至第二次逮捕的时间作为变量标签。
- 单击**完成**。

运行分析

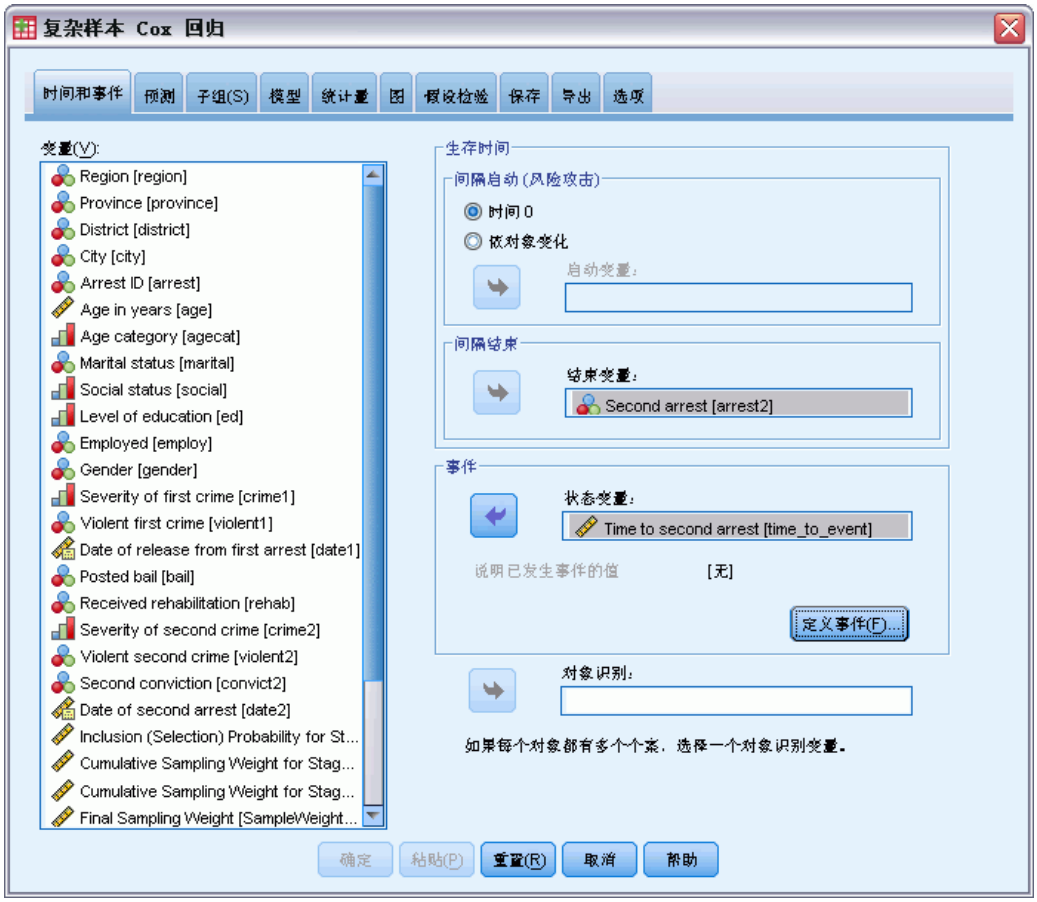
- 要运行复杂样本 Cox 回归分析，请从菜单中选择：
  分析 > **复杂样本** > **Cox 回归...**

图片 22-7
“Cox 回归的复杂样本计划”对话框



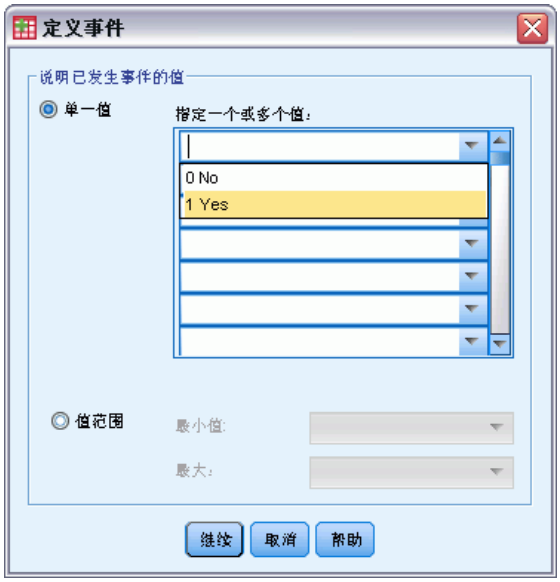
- 浏览至样本文件目录并选择 recidivism_cs.csplan 作为计划文件。
- 在“联合概率”组中选择**定制文件**，浏览至样本文件目录，并选择 recidivism_cs_jointprob.sav。
- 单击**继续**。

图片 22-8
“Cox 回归”对话框，“时间与事件”选项卡



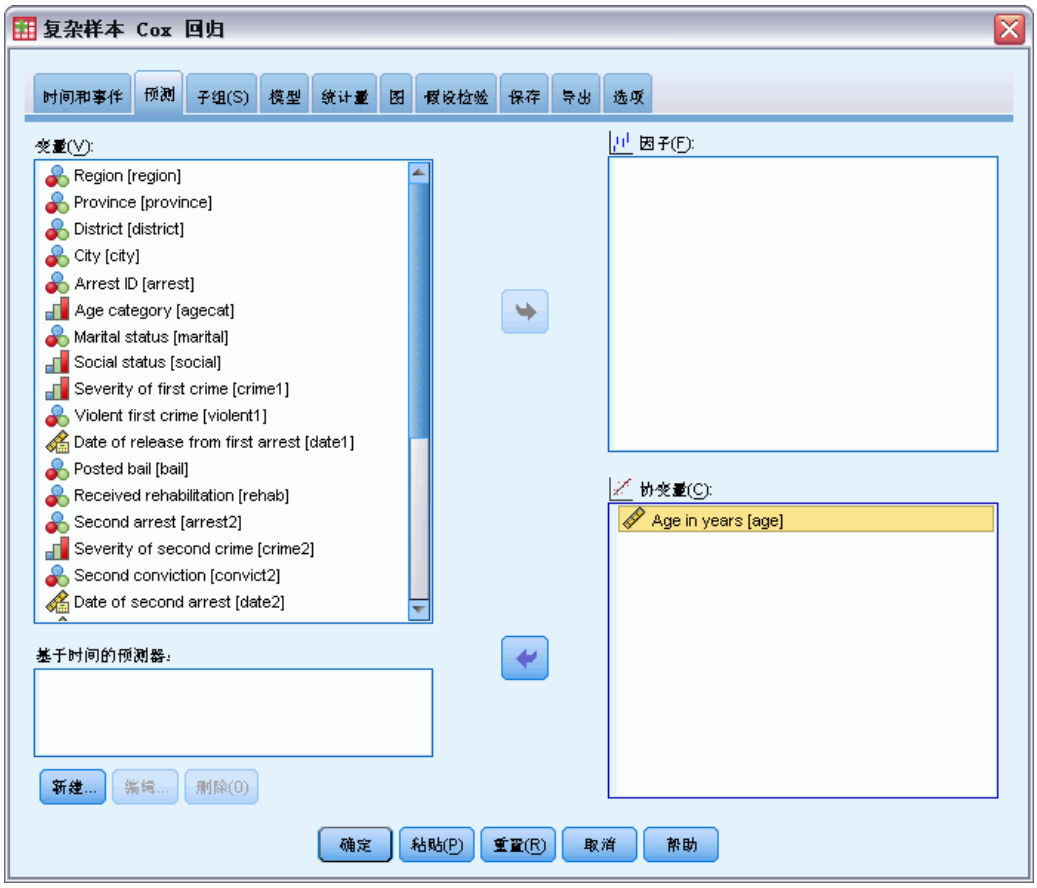
- 选择至第二次逮捕的时间 [time_to_event] 作为界定区间结束点的变量。
- 选择第二次逮捕 [arrest2] 作为界定事件是否发生的变量。
- 单击**定义事件**。

图片 22-9
“定义事件”对话框



- 选择 1 是 作为指示感兴趣事件（重新逮捕）已经发生的值。
- 单击继续。
- 单击预测变量选项卡。

图片 22-10
“Cox 回归”对话框，“预测器”选项卡



- 选择 年龄 [age] 作为协变量。
- 单击**统计量**选项卡。

图片 22-11
“Cox 回归”对话框，“统计量”选项卡

- 选择比例危险测试然后选择对数作为“模型假设”组中的时间函数。
- 选择其他模型的参数估计值。
- 单击确定。

样本设计信息

图片 22-12
样本设计信息

			N
未加权的计数	有效	被试变量	5687
		案例	5687
		无效个案	0
		总个案数	5687
有效		种群被试变量大小	307583.898
阶段 1	有效	分层	4
		单位	20
有效		抽样设计的自由度	16

此表包含与模型估计相关的样本设计信息。

- 每个主体有一个个案，且所有 5,687 个个案都用于分析。
- 样本表示少于整个估计总体的 2%。
- 在设计的第一步中，设计需要 4 个层次和每层次 5 个单元,总计 20 个单元。抽样设计自由度估计为 20−4=16。

模型效应检验

图片 22-13
模型效应检验

源	df1	df2	Wald F	Sig.
age	1.000	16.000	504.787	1.580E-13

生存时间变量: Time to second arrest
事件状态变量: Second arrest = 1
模型: age

在成比例风险模型中，预测器年龄的显著性值小于 0.05，所以看起来对模型有贡献。

比例危险测试

图片 22-14
比例危险完整测试

df1	df2	Wald F	Sig.
1.000	16.000	29.924	5.136E-5

生存时间变量: Time to second arrest
事件状态变量: Second arrest = 1
模型: age, age*_TF

图片 22-15
其他模型的参数估计值

			95% 置信区间	
参数	B	标准 误差	下限	下限
age	-.002	.014	-.025	0.02
age*_TF[a]	-.012	.002	-.016	-.009

生存时间变量: Time to second arrest
事件状态变量: Second arrest = 1
模型: age, age*_TF

a. 时间函数: Log

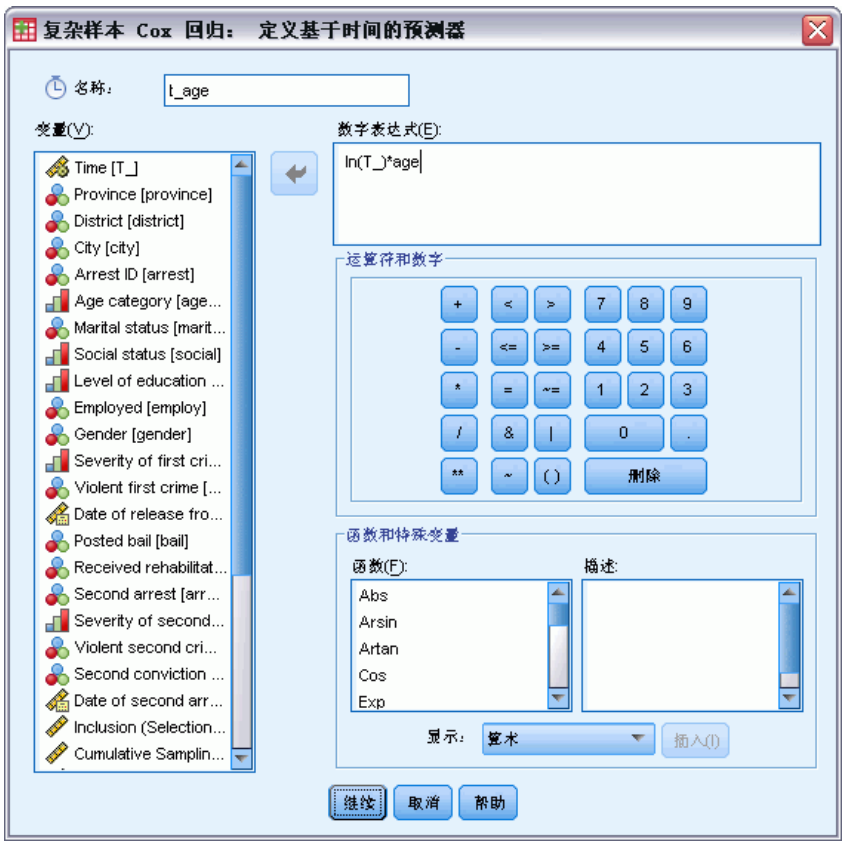
比例危险完整测试的显著性值小于 0.05，这表示违反了比例危险假设。对数时间函数用于其他模型，所以很容易复制此依时预测器。

添加依时预测器

▶ 调用“复杂样本 Cox 回归”对话框并单击**预测器**选项卡。

▶ 单击新建。

图片 22-16
“Cox 回归界定依时预测器”对话框

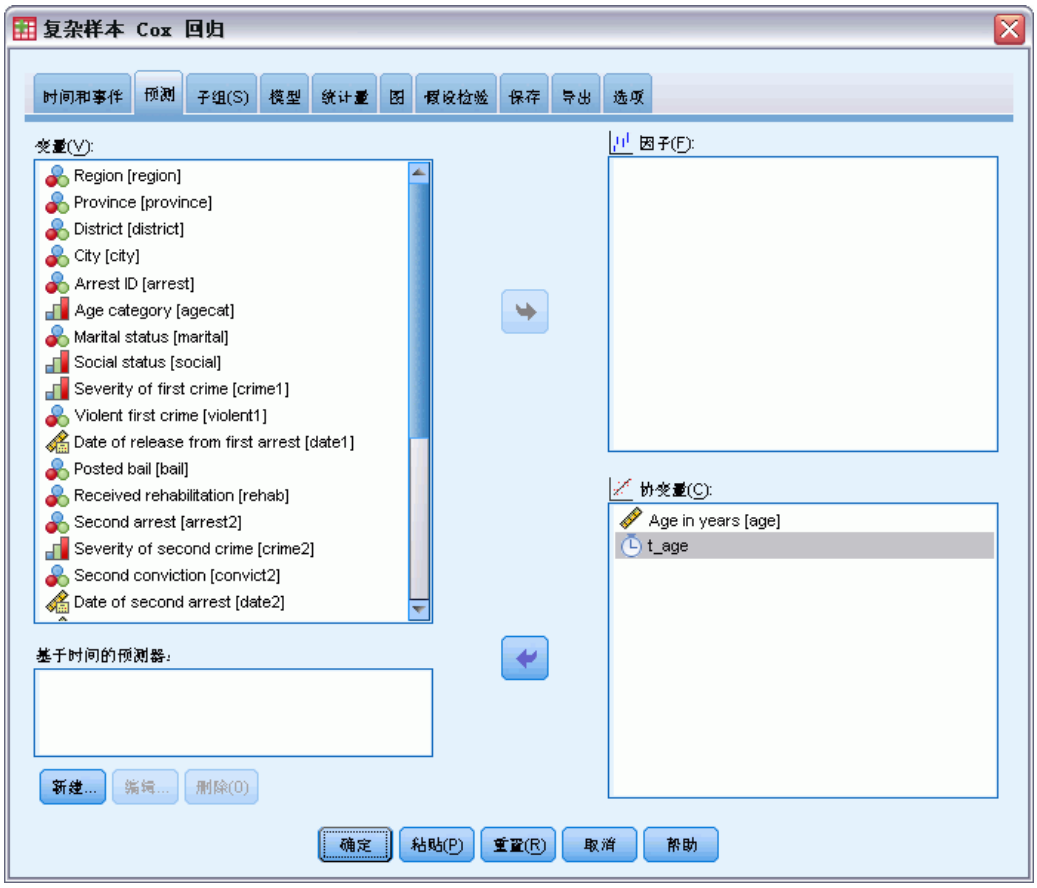


▶ 键入 t_age 作为您希望定义的依时预测器的名称。

▶ 键入 ln(T_)*age 作为数值表达式。

▶ 单击继续。

图片 22-17
“Cox 回归”对话框，“预测器”选项卡



- 选择 t_age 作为协变量。
- 单击**统计量**选项卡。

图片 22-18
“Cox 回归”对话框，“预测器”选项卡

- 选择“参数”组中的估算、标准误、置信区间以及设计效应。
- 取消选择“模型假设”组中的比例危险测试和其他模型的参数估计值。
- 单击确定。

模型效应检验

图片 22-19
模型效应检验

源	df1	df2	Wald F	Sig.
age	1.000	16.000	.015	.905
t_age	1.000	16.000	29.924	5.136E-5

生存时间变量: Time to second arrest
事件状态变量: Second arrest = 1
模型: age, t_age

添加了依时预测器后，年龄的显著性值将变为 0.91，表示其对模型的贡献将由 t_age 的贡献所取代。

参数估计值

图片 22-20
参数估计值

参数	B	标准 误差	95% 置信区间 下限	下限	设计效果
age	-.002	.014	-.030	.027	.702
t_age	-.012	.002	-.017	-.008	.666

生存时间变量: Time to second arrest
事件状态变量: Second arrest = 1
模型: age, t_age

请看参数估计值和标准误，您可以看到您已经从比例危险测试中复制了其他模型。通过明确指定模型，您可以请求附加的参数统计量和图。此处，我们已经请求了设计效应；如果您假设数据集为简单的随机样本，那么 t_age 的值小于 1 则表示 t_age 的标准误小于您将得到的值。在这种情况下，t_age 的效应仍将具有统计显著性，但是置信区间会更宽。

“复杂样本 Cox 回归”中的每个主体多个个案

正在研究结束缺血性中风后复元计划的患者存活时间的研究人员面临着很多挑战。

每个主体多个个案。代表患者医疗历史的变量和预测器一样有用。随着时间变化，患者可能会经历改变其医疗历史的重要医疗事件。在此数据集中，记录了心肌梗塞、缺血性中风或出血性中风的发生和记录事件的时间。您可以在过程中创建可计算依时协变量以便将此信息包括在模型中，但是如果使用每个主体多个个案将会更方便。请注意，变量为原始编码以便通过变量记录患者历史，因此您需要重组数据集。

左侧截短。缺血性中风的风险攻击开始。但是，样本只包含复元计划中存活的患者，就观察到的存活时间被复元长度“夸大”而言，则样本为左侧截短。可以通过将他们结束复元的时间指定为进入研究的时间对此进行说明。

无抽样计划。数据集未通过复杂抽样计划收集，且被视为简单随机样本。您需要创建一个分析计划来使用“复杂样本 Cox 回归”。

该数据集收集在 stroke_survival.sav 中。有关详细信息，请参阅第 251 页码附录 A 中的样本文件.使用“重组数据向导”为分析准备数据，然后使用“分析准备向导”创建简单随机抽样计划，最后使用“复杂样本 Cox 回归”为存活时间建模。

准备数据以进行分析

重组数据前，您需要创建两个辅助变量以帮助重组。

- 要计算新变量，请从菜单中选择：
  转换 > 计算变量...

图片 22-21
“计算变量”对话框

- 键入 start_time2 作为目标变量。
- 键入 time1 作为数值表达式。
- 单击**确定**。

▶ 调用“计算变量”对话框。

图片 22-22
“计算变量”对话框

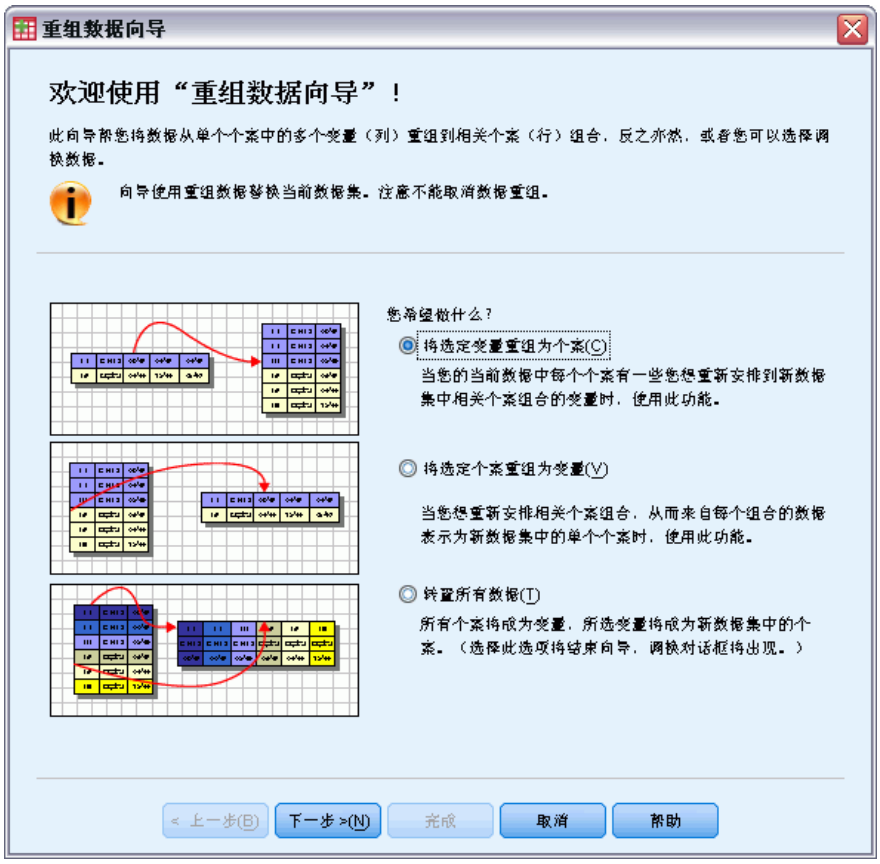
▶ 键入 start_time3 作为目标变量。

▶ 键入 time2 作为数值表达式。

▶ 单击**确定**。

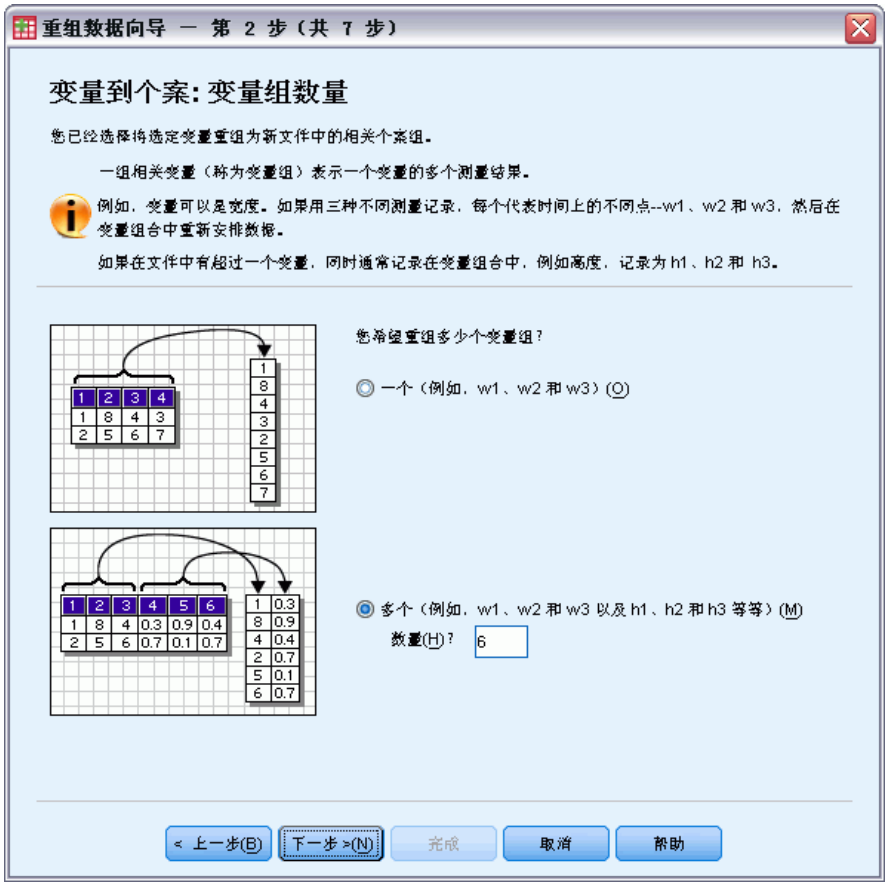
▶ 要从个案变量中重组数据，请从菜单中选择：
数据 > **重组**...

图片 22-23
重组数据向导，“欢迎”步骤



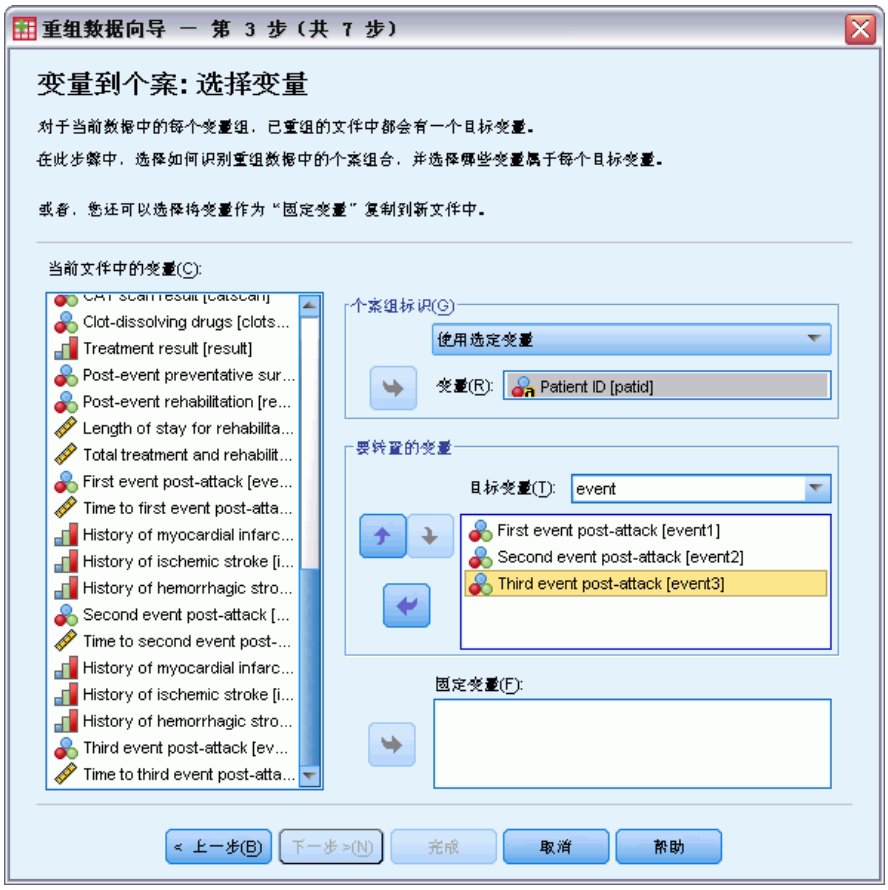
- 确保选择了**将选定变量重组为个案**。
- 单击**下一步**。

图片 22-24
重组数据向导，“变量组的变量到个案数”步骤



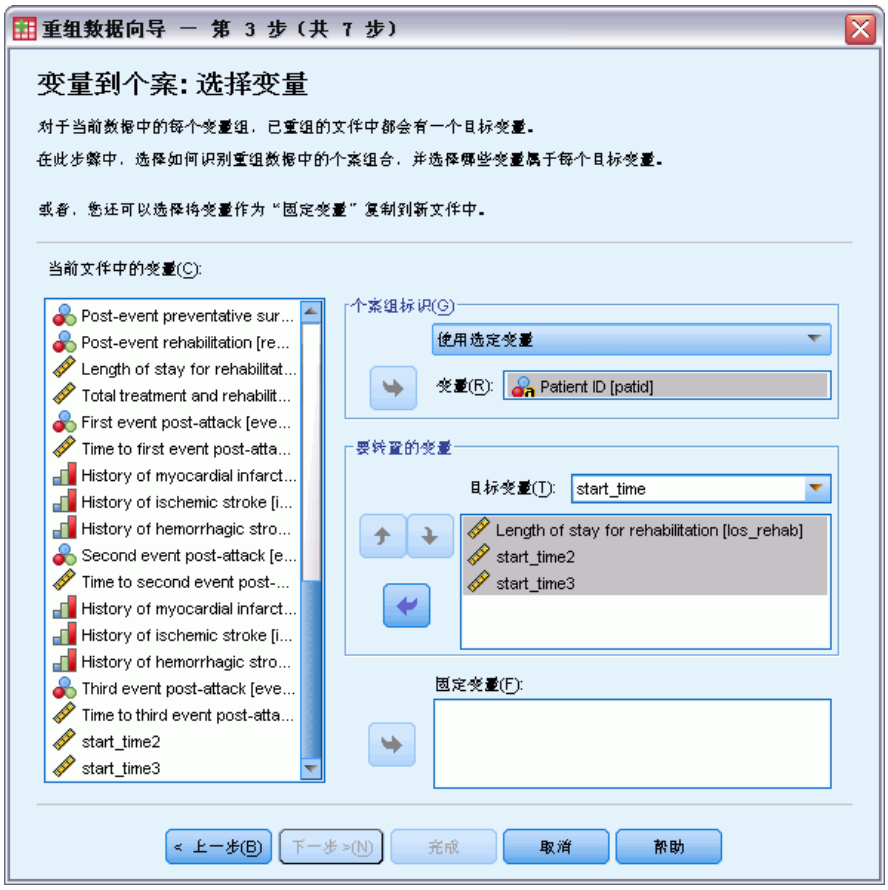
- 选择多个变量组进行重组。
- 键入 6 作为组数目。
- 单击下一步。

图片 22-25
重组数据向导，“变量到个案选择变量”步骤



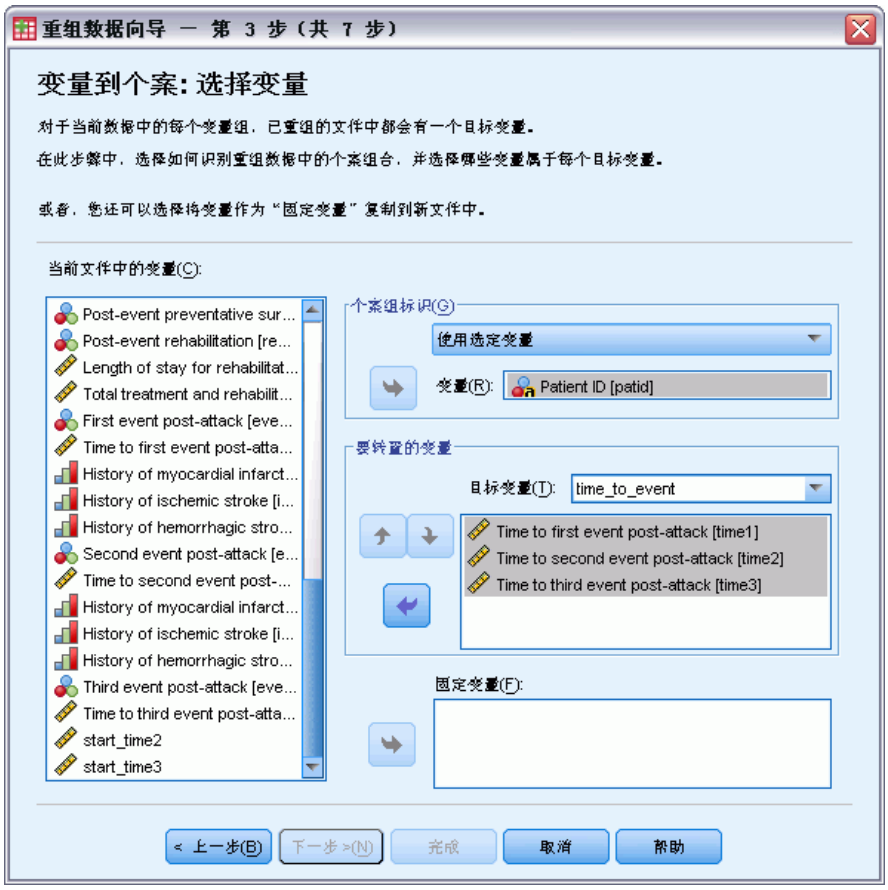
- 在“个案组标识”组中，选择**使用选择的变量**并选择 Patient ID [patid] 作为主体标识。
- 键入 **event** 作为第一个目标变量。
- 选择第一个事件后攻击 [event1]、第二个事件后攻击 [event2] 和 第三个事件后攻击 [event3] 作为即将转置的变量。
- 从目标变量列表中选择 trans2。

图片 22-26
重组数据向导，“变量到个案选择变量”步骤



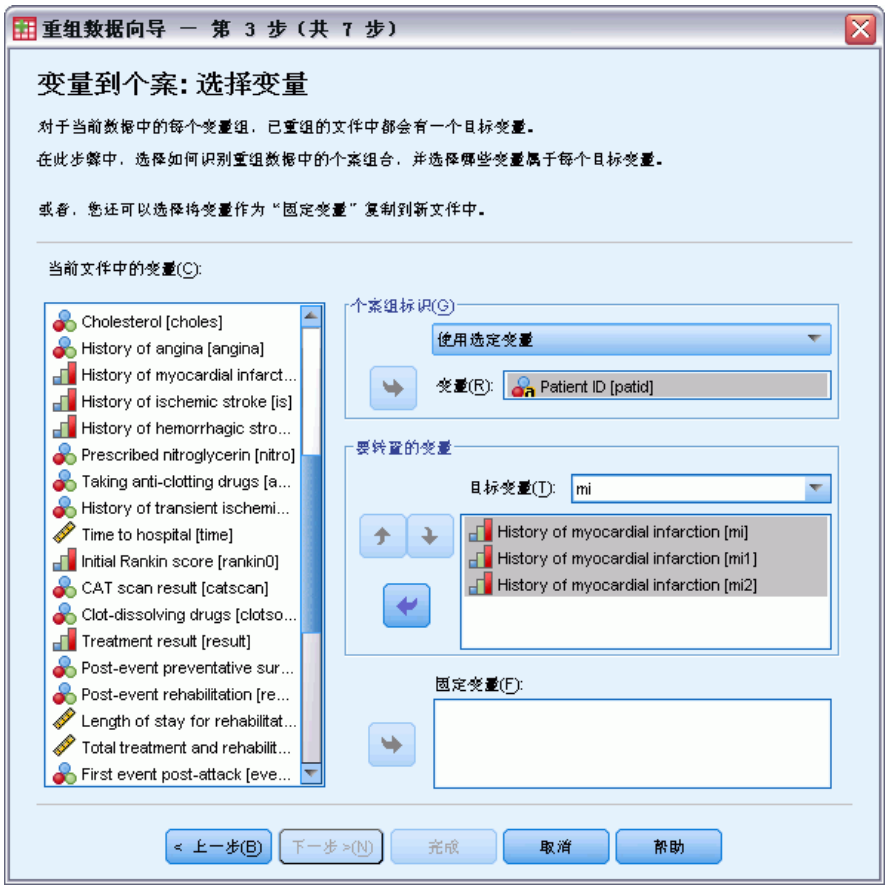
- 键入 start_time 作为目标变量。
- 选择住院复元时间 [los_rehab]、start_time2、和 start_time3 作为即将转置的变量。第一个事件后攻击时间 [time1] 和 第二个事件后攻击时间 [time2] 将用于创建结束时间，且每个变量只可以出现在一个即将转置的变量列表中，这样就需要 start_time2 和 start_time3。
- 从目标变量列表中选择 trans3。

图片 22-27
重组数据向导，“变量到个案选择变量”步骤



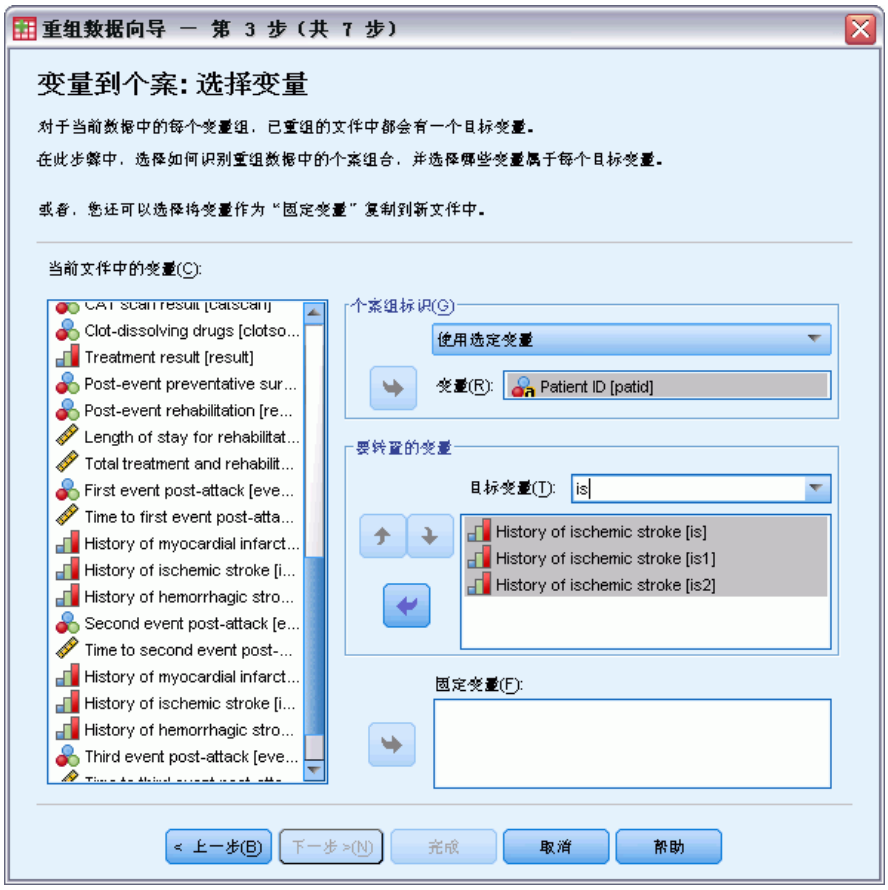
- 键入 time_to_event 作为目标变量。
- 选择第一个事件后攻击时间 [time1]、第二个事件后攻击时间 [time2] 和 第三个事件后攻击时间 [time3] 作为即将转置的变量。
- 从目标变量列表中选择 trans4。

图片 22-28
重组数据向导，“变量到个案选择变量”步骤



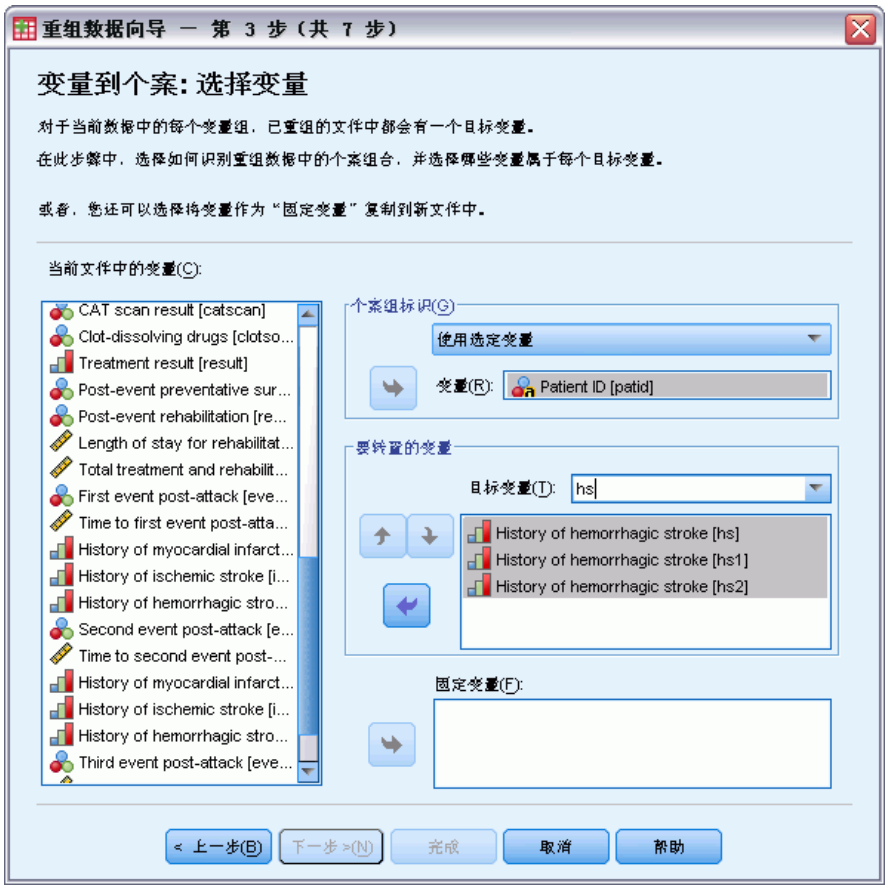
- 键入 mi 作为目标变量。
- 选择心肌梗塞历史 [mi]、心肌梗塞历史 [mi1]、和心肌梗塞历史 [mi2] 作为即将转置的变量。
- 从目标变量列表中选择 trans5。

图片 22-29
重组数据向导，“变量到个案选择变量”步骤



- 键入 is 作为目标变量。
- 选择缺血性中风历史 [is]、缺血性中风历史 [is1]、和缺血性中风历史 [is2] 作为即将转置的变量。
- 从目标变量列表中选择 trans6。

图片 22-30
重组数据向导，“变量到个案选择变量”步骤



- 键入 hs 作为目标变量。
- 选择出血性中风历史 [hs]、出血性中风历史 [hs1]、和出血性中风历史 [hs2] 作为即将转置的变量。
- 单击**下一步**，然后在“创建索引变量”步骤中单击**下一步**。

图片 22-31
重组数据向导，“变量到个案创建索引变量”步骤

- 键入 event_index 作为索引变量的名称并键入 **事件索引** 作为变量标签。
- 单击**下一步**。

图片 22-32
重组数据向导，“变量到个案创建索引变量”步骤

- 确保选择了保留并视为固定变量。
- 单击完成。

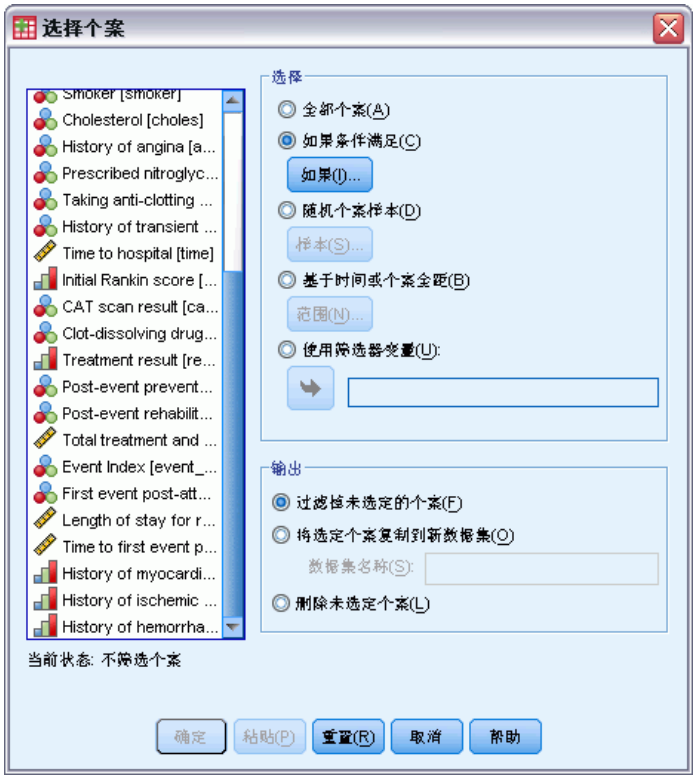
图片 22-33
重组的数据

event_index	event	start_time	time_to_event	mi	is	hs
1	0	3	1500	0	1	0
2	-4	1500	-4	-4	-4	-4
3	-4	.	-4	-4	-4	-4
1	1	33	1311	0	1	0
2	4	1311	1325	1	1	0
3	-3	1325	-3	-3	-3	-3
1	4	12	1098	1	1	0
2	-3	1098	-3	-3	-3	-3
3	-3	.	-3	-3	-3	-3
1	4	4	1356	0	1	0
2	-3	1356	-3	-3	-3	-3
3	-3	.	-3	-3	-3	-3

重组过的数据包含每个患者的三个个案；但是，许多患者所经历过的事件少于三个，因此有很多具有事件负（缺失）值的个案。您可以将这些个案从数据集中过滤掉。

- 要过滤这些个案，请从菜单中选择：
  数据 > 选择个案...

图片 22-34
“选择个案”对话框

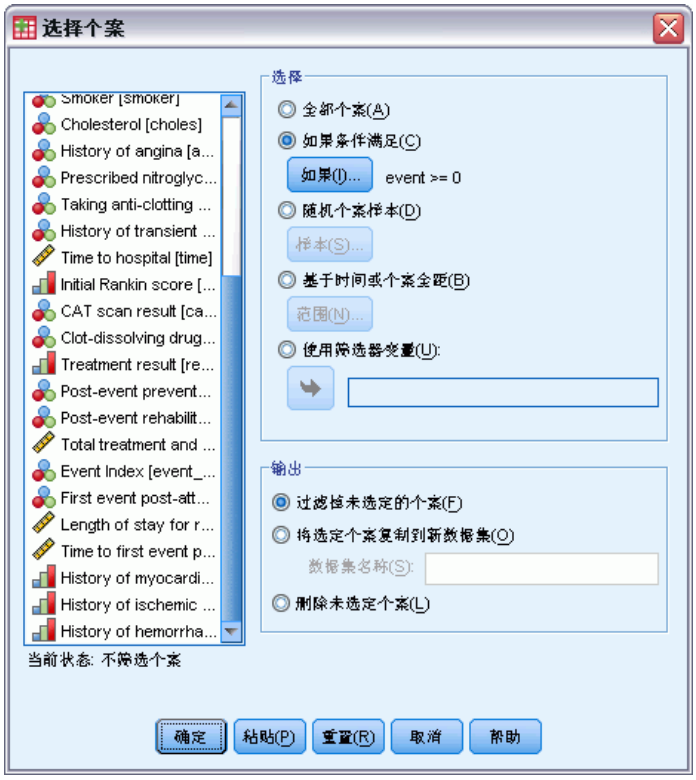


- 选择如果满足条件。
- 单击如果。

图片 22-35
“选择个案：If”对话框

- 键入 event >= 0 作为条件表达式。
- 单击继续。

图片 22-36
“选择个案”对话框



- 选择**删除未选定个案**。
- 单击**确定**。

创建简单的随机抽样分析计划

现在即可创建简单的随机抽样分析计划。

- 首先，您需要创建一个抽样权重变量。从菜单中选择：
  转换 > **计算变量...**

图片 22-37
“Cox 回归”主对话框

- 键入 sampleweight 作为目标变量。
- 键入 1 作为数值表达式。
- 单击**确定**。

  现在即可创建分析计划。

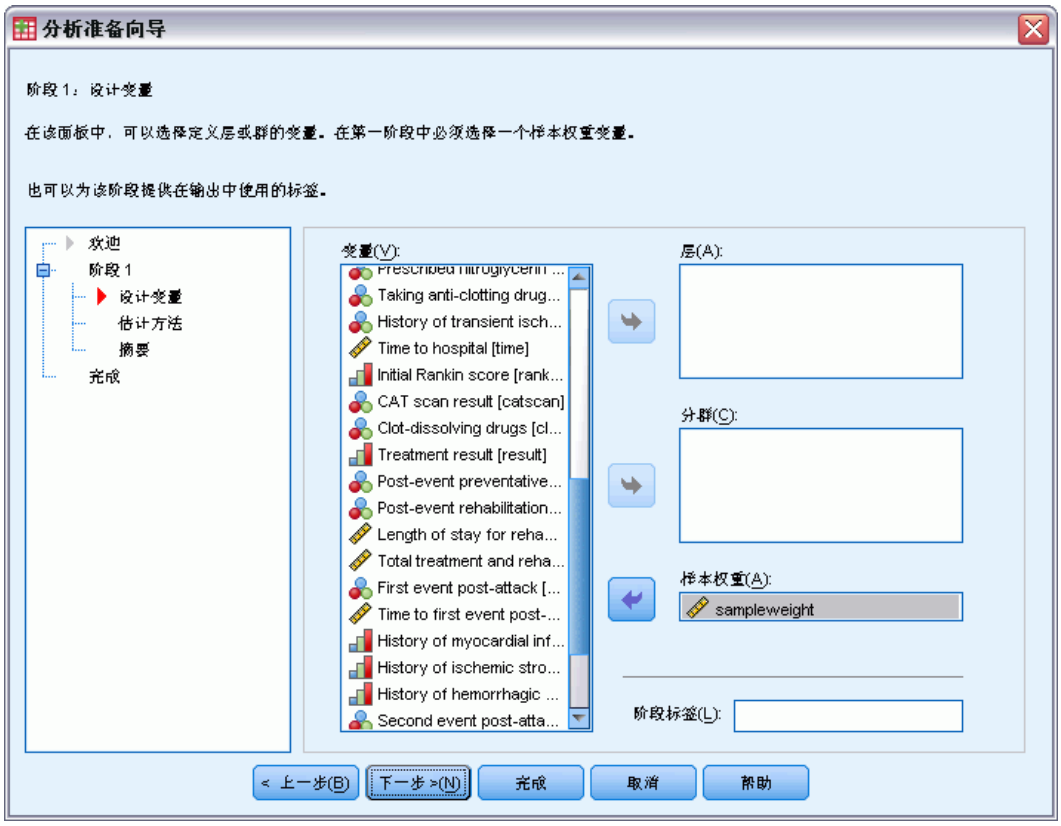
  注意：如果您想跳过以下说明并继续数据分析，您可以使用样本文件目录中已有的计划文件 srs.csaplan。

- 要创建分析计划，请从菜单中选择：
  分析 > 复杂样本 > 准备分析...

图片 22-38
分析准备向导，“欢迎”步骤

- 选择**创建计划文件**并键入 srs.csaplan 作为文件名。此外，浏览至您想保存的位置。
- 单击**下一步**。

图片 22-39
分析准备向导，“设计变量”



- 选择 sampleweight 作为样本权重变量。
- 单击下一步。

图片 22-40
分析准备向导，“估计方法”

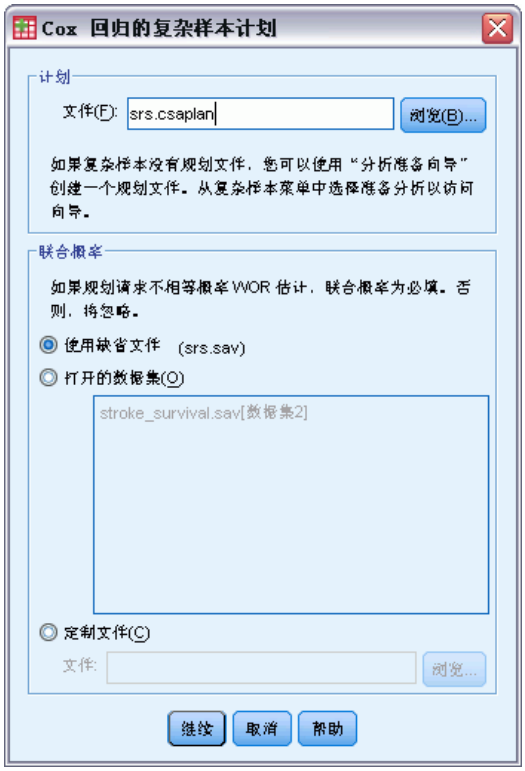
- 取消选择**使用有限总体纠正**。
- 单击**完成**。

现在即可运行分析。

运行分析

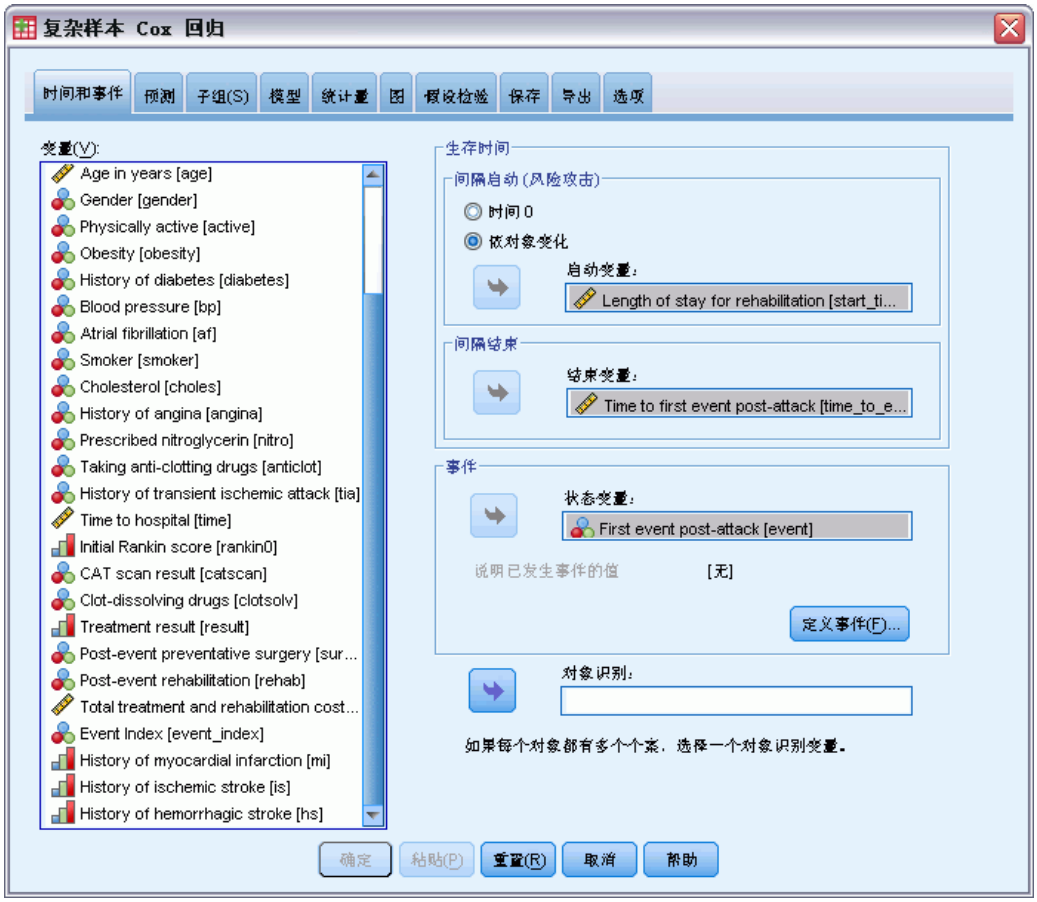
- 要运行复杂样本 Cox 回归分析，请从菜单中选择：
  分析 > 复杂样本 > Cox 回归...

图片 22-41
“Cox 回归计划”对话框



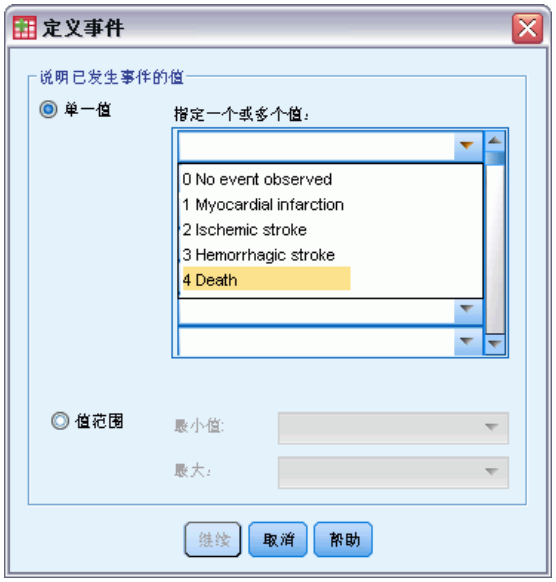
- 浏览至您保存简单随机抽样分析计划的位置，或样本文件目录，并选择 srs.csaplan。
- 单击继续。

图片 22-42
“Cox 回归”对话框，“时间与事件”选项卡



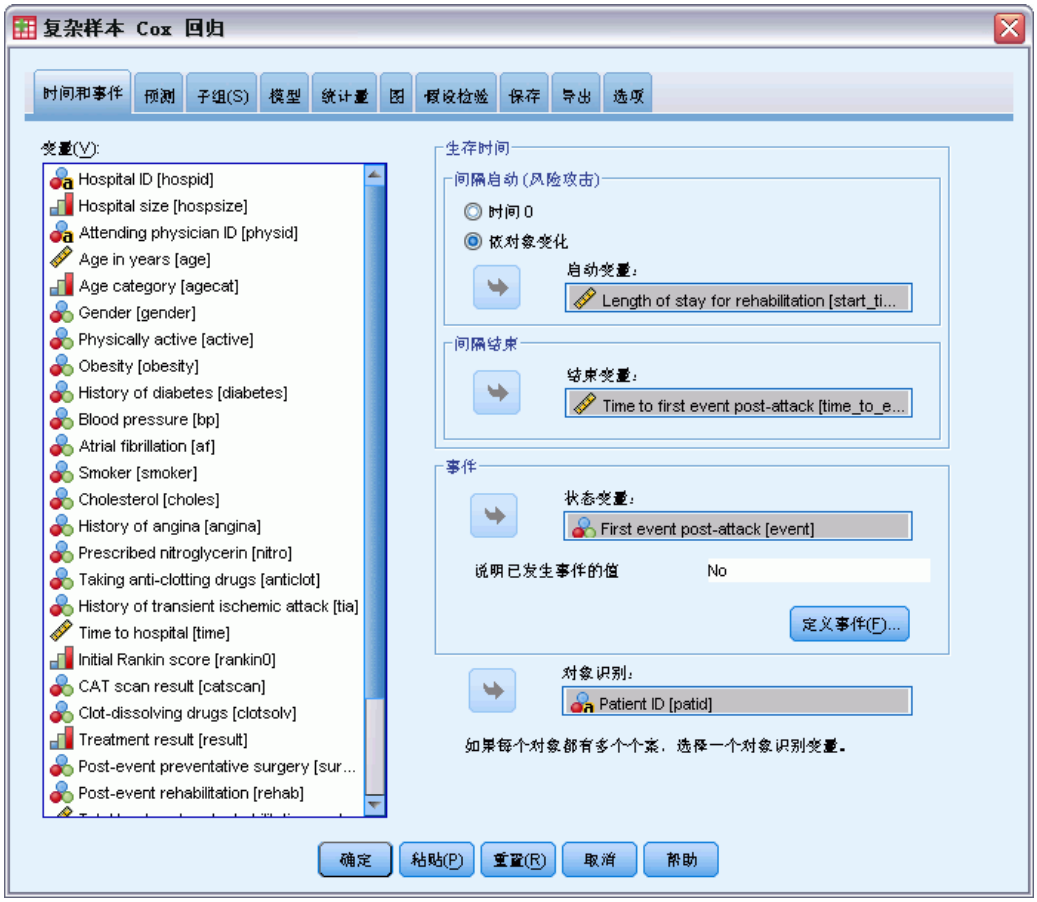
- 选择**按主体变化**并选择住院复元时间 [los_rehab] 作为开始变量。请注意，重组的变量使用了用于构建它的第一个变量中的变量标签，尽管此标签并不很适合构建的变量。
- 选择第一个事件后攻击时间 [time_to_event] 作为结束变量。
- 选择第一个事件后攻击 [event] 作为状态变量。
- 单击**定义事件**。

图片 22-43
“定义事件”对话框



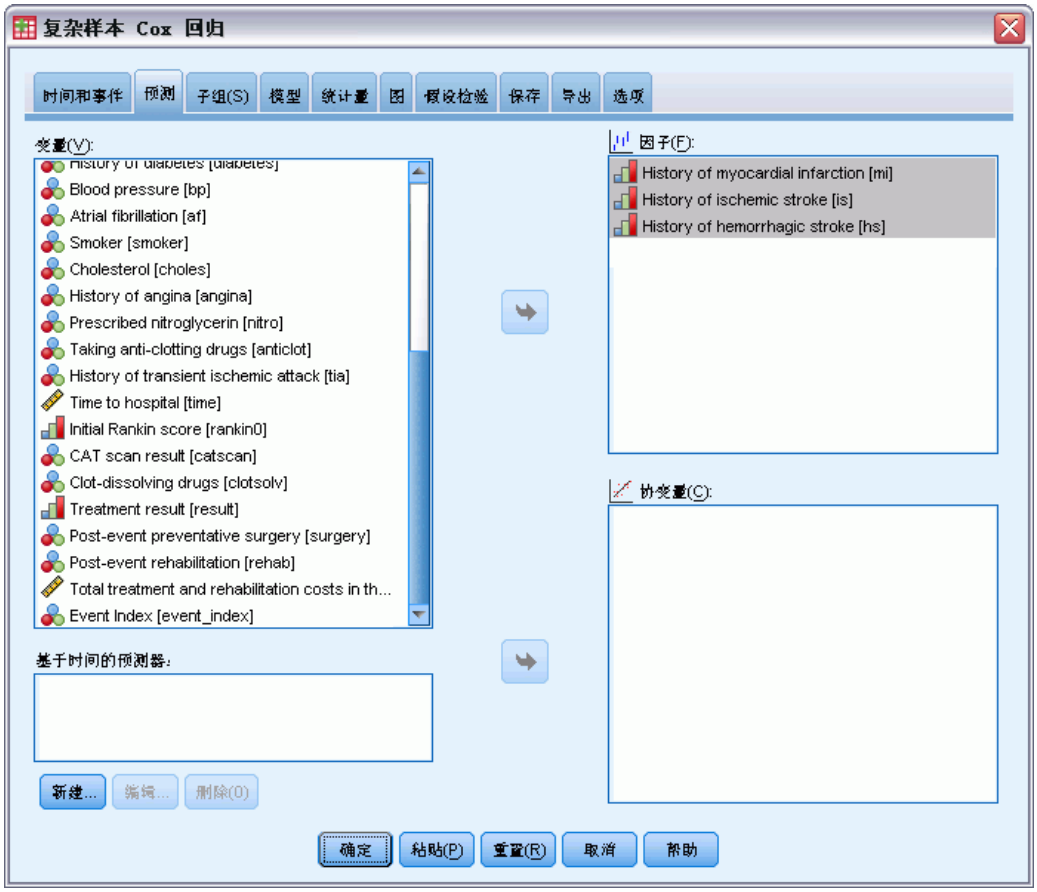
- 选择 4 死亡 作为指示终端事件已经发生的值。
- 单击继续。

图片 22-44
“Cox 回归”对话框，“时间与事件”选项卡



- 选择 Patient ID [patid] 作为主体标识。
- 单击**预测变量**选项卡。

图片 22-45
“Cox 回归”对话框，“预测器”选项卡

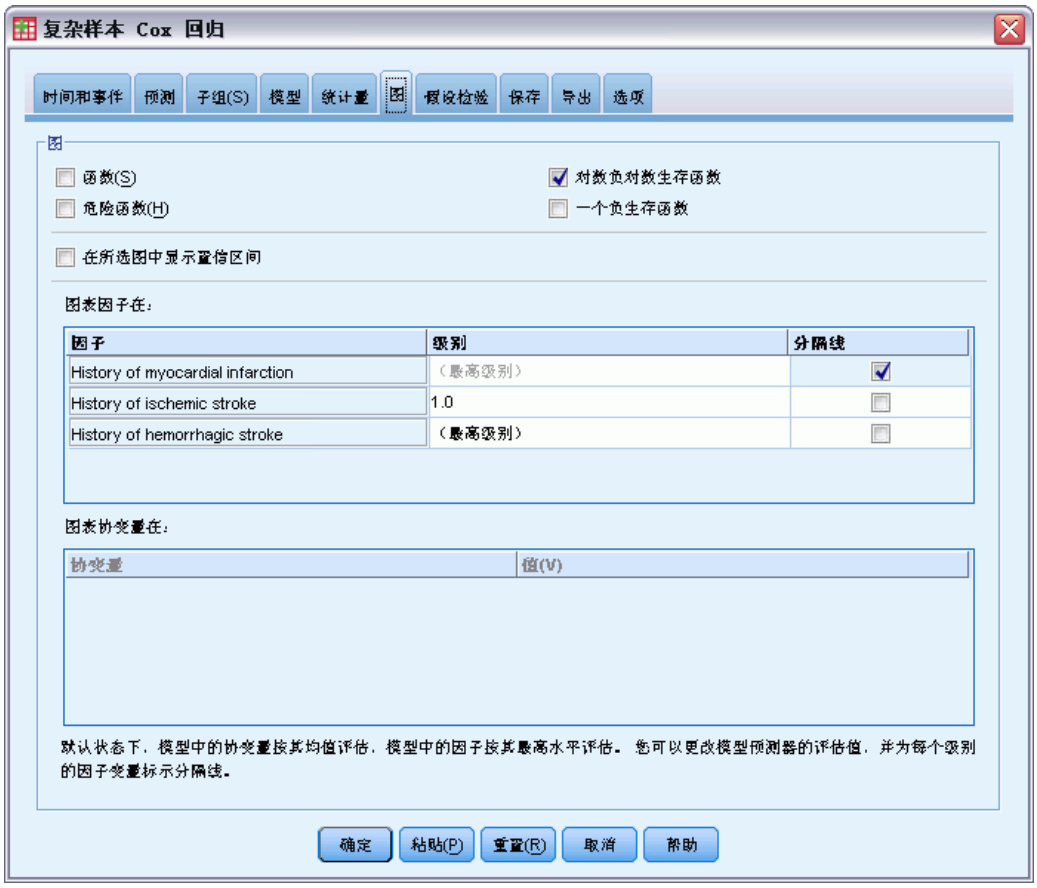


- 选择心肌梗塞历史 [mi] 到 出血性中风历史 [hs] 作为因子。
- 单击**统计量**选项卡。

图片 22-46
“Cox 回归”对话框，“统计量”选项卡

- 选择“参数”组中的**估算**、**取幂估值**、**标准误**和**置信区间**。
- 单击**图**选项卡。

图片 22-47
“Cox 回归”对话框，“统计量”选项卡



- 选择**对数负对数生存函数**。
- 检查心肌梗塞历史的**分隔线**。
- 选择 1.0 作为缺血性中风历史的水平。
- 选择 0.0 作为出血性中风历史的水平。
- 单击**选项**选项卡。

图片 22-48
“Cox 回归”，“选项”选项卡

- 在“估计”组中选择 Breslow 作为断开连接方法。
- 单击确定。

样本设计信息

图片 22-49
样本设计信息

			N
未加权的计数	有效	被试变量	2421
		案例	3310
		无效个案	0
		总个案数	3310
有效		种群被试变量大小	2421.000
阶段 1	有效	分层	1
		单位	2421
有效		抽样设计的自由度	2420

此表包含与模型估计相关的样本设计信息。

- 一些主体有多个个案，且所有 3,310 个个案都用于分析。
- 设计具有单个分层和 2,421 个单位（每个主题一个）。抽样设计自由度估计为 2421−1=2420。

模型效应检验

图片 22-50
模型效应检验

源	df1	df2	Wald F	Sig.
mi	3.000	2418.000	452.873	.000
is	2.000	2419.000	1064.936	.000
hs	2.000	2419.000	739.197	.000

生存时间变量: Length of stay for rehabilitation, Time to first event post-attack
事件状态变量: First event post-attack = 4
被试 ID 变量: Patient ID
模型: mi, is, hs

每种效应的显著性值接近于 0，说明它们对模型都有贡献。

参数估计值

图片 22-51
参数估计值

			95% 置信区间			Exp(B) 的 95% 置信区间	
参数	B	标准 误差	下限	下限	Exp(B)	下限	下限
[mi=0]	-6.381	.283	-6.935	-5.827	.002	.001	.003
[mi=1]	-5.589	.284	-6.147	-5.032	.004	.002	.007
[mi=2]	-2.119	.344	-2.794	-1.445	.120	.061	.236
[mi=3]	.000[a]	.	.	.	1.000	.	.
[is=1]	-6.421	.202	-6.817	-6.024	.002	.001	.002
[is=2]	-2.803	.222	-3.239	-2.366	.061	.039	.094
[is=3]	.000[a]	.	.	.	1.000	.	.
[hs=0]	-6.148	.355	-6.844	-5.453	.002	.001	.004
[hs=1]	-2.232	.373	-2.963	-1.502	.107	.052	.223
[hs=2]	.000[a]	.	.	.	1.000	.	.

生存时间变量: Length of stay for rehabilitation, Time to first event post-attack
事件状态变量: First event post-attack = 4
被试 ID 变量: Patient ID
模型: mi, is, hs

a. 设置为零，原因是此参数为冗余的。

b. 断开连接方法: Breslow

过程使用每个因子的最后类别作为参考类别；其他类别的效应相对于参考类别。请注意，当估计值可用于统计量测试时，取幂估值，即 Exp(B) 更容易解释为相对于参考类别的风险中的预测变化。

- [mi=0] 的 Exp(B) 值意味着之前未患过心肌梗塞 (mi) 的患者的死亡风险是之前患过心肌梗塞 (mi) 患者的 0.002 倍。

- [mi=1] 和 [mi=0] 的置信区间重叠，这表示之前患过一次心肌梗塞 (mi) 患者的死亡风险与之前未患过心肌梗塞 (mi) 的患者并无统计上的区别。
- [mi=0] 和 [mi=1] 的置信区间未和 [mi=2] 的区间重叠，且任何一个都不包括 0。所以，看起来那些未患过或患过一次心肌梗塞 (mi) 的患者的风险和患过两次心肌梗塞 (mi) 的患者的风险是有区别的，同样，和患过三次心肌梗塞 (mi) 的患者的风险也是有区别的。

类似的关系也适用于 is 和 hs 的水平，即之前患病次数增加，死亡的风险也随之增加。

模式值

图片 22-52
模式值

		生存时间间隔				
		启动	结束	History of myocardial infarction	History of ischemic stroke	History of hemorrhagic stroke
参考图案	1	.000	[a]	Three	Three	Two
图案 1.1	1	.000	[a]	None	One	None
图案 1.2	1	.000	[a]	One	One	None
图案 1.3	1	.000	[a]	Two	One	None
图案 1.4	1	.000	[a]	Three	One	None

参考图案中此预测器的值未指定预测器。
每个生存时间间隔被定义为启动 < 生存时间 <= 结束。
模型: mi, is, hs.

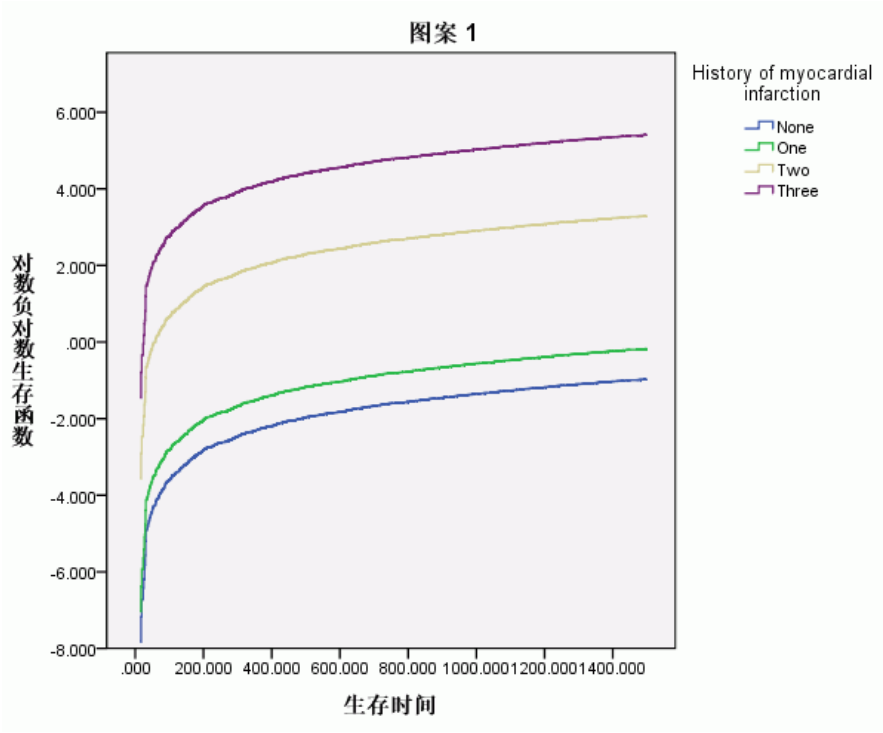
a. 未限制

模式值表列出了定义每个预测器模式的值。除了模型中的预测器之外，还显示了生存区间的开始和结束时间。对于从对话框运行分析，开始和结束时间始终分别为 0 和无限制；通过语法您可以指定分段恒定预测器路径。

- 参考模式在每个因子的参考类别和每个协变量的均值处设置（此模型中没有协变量）。对于此数据集来说，不可能发生参考模型显示的因子组合，所以我们将忽略参考模式的对数负对数图。
- 模式 1.1 到 1.4 仅在心肌梗塞历史的值上有所不同。在其他变量保持恒定的同时，为心肌梗塞历史的每个值创建分隔模式（及要求的图中的分隔线）。

对数负对数图

图片 22-53
对数负对数图



此图显示生存函数的对数负对数，ln(−ln(suvival))，和生存时间。这一特殊的图显示心肌梗塞历史的每个类别的分隔曲线，其中缺血性中风历史设定为一而出血性中风历史设定为无，且此图是生存函数上心肌梗塞历史的效应的有用可视表达。正如在参数估计值表中看到的那样，那些未患过或患过一次心肌梗塞的患者的生存和患过两次心肌梗塞的患者的生存是有区别的，同样，和患过三次心肌梗塞的患者的生存也是有区别的。

摘要

您已经为后攻击生存拟合了 Cox 回归模型，这些后攻击生存会估计更改后攻击患者历史的效应。这仅仅是个开始，因为研究人员毫无疑问想将其他潜在的预测器包含进模型中。除此以外，在对此数据集的进一步分析中，您可能要考虑更多模型构造上的显著变化。例如，当前模型假设患者历史变化事件的效应可以按照基线风险乘数来进行量化。相反，假设基线风险的形状由非死亡事件的发生所改变也是合理的。要完成这些，您可以对基于事件索引的分析进行分层。

附录
A

样本文件

随产品一起安装的样本文件可以在安装目录的 Samples 子目录中找到。对于以下每种语言在“样本”子目录中有单独的文件夹：英语、法语、德语、意大利语、日语、韩语、波兰语、俄语、简体中文、西班牙语和繁体中文。

并非所有样本文件均提供此处的全部语言版本。如果样本文件未提供某种语言的版本，则相应语言文件夹中包含该样本文件的英语版本。

描述

以下是对在整个文档的各种示例中使用的样本文件的简要描述。

- **accidents.sav。** 该假设数据文件涉及某保险公司，该公司正在研究给定区域内汽车事故的年龄和性别风险因子。每个个案对应一个年龄类别和性别类别的交叉分类。
- **adl.sav。**该假设数据文件涉及在确定针对脑卒中患者的建议治疗类型的优点方面的举措。医师将女性脑卒中患者随机分配到两组中的一组。第一组患者接受标准的物理治疗，而第二组患者则接受附加的情绪治疗。在进行治疗的三个月时间里，将为每个患者进行一般日常生活行为的能力评分并作为原始变量。
- **advert.sav。**该假设数据文件涉及某零售商在检查广告支出与销售业绩之间的关系方面的举措。为此，他们收集了过去的销售数据以及相关的广告成本..
- **aflatoxin.sav。**该假设数据文件涉及对谷物的黄曲霉毒素的检测，该毒素的浓度会因谷物产量的不同（不同谷物之间及同种谷物之间）而有较大变化。谷物加工机从 8 个谷物产量的每一个中收到 16 个样本并以十亿分之几 (PPB) 为单位来测量黄曲霉毒素的水平。
- **aflatoxin20.sav。**该数据文件包括对数据文件 aflatoxin.sav 中产量 4 和 8 的 16 个样本中的每一个样本进行的黄曲霉毒素度量。
- **anorectic.sav。**在研究厌食/暴食行为的标准症状参照时，研究人员(Van der Ham, Meulman, Van Strien, 和 Van Engeland, 1997)对 55 名已知存在进食障碍的青少年进行了调查。其中每名患者每年都将进行四次检查，因此总观测数为 220。在每次观测期间，将对这些患者按 16 种症状逐项评分。但 71 号和 76 号患者的症状得分均在时间点 2 缺失，47 号患者的症状得分在时间点 3 缺失，因此有效观测数为 217。
- **autoaccidents.sav。**该假设数据文件涉及某保险分析师在为每个驾驶员的汽车事故数量建模方面的举措，同时也解释了驾驶员年龄和性别与汽车事故数量之间的关系。每个个案代表单独的驾驶员并记录驾驶员的性别、年龄以及最近五年内的汽车事故数量。
- **band.sav。**该数据文件包含某乐队音乐 CD 的假设每周销售数据。还包括三个可能的预测变量的数据。

- **bankloan.sav。**该假设数据文件涉及某银行在降低贷款拖欠率方面的举措。该文件包含 850 位过去和潜在客户的财务和人口统计信息。前 700 个个案是以前曾获得贷款的客户。剩下的 150 个个案是潜在客户，银行需要按高或低信用风险对他进行分类。
- **bankloan_binning.sav。**该假设数据文件包含 5,000 位过去客户的财务和人口统计信息。
- **behavior.sav。**在一个经典示例中(Price 和 Bouffard, 1974)，52 名学生被要求以 10 分的标度对 15 种情况和 15 种行为的组合进行评价，该 10 分的标度介于 0 =平均值在个人值之上，值被视为相异性。
- **behavior_ini.sav。**该数据文件包含 behavior.sav 的二维解的初始配置。
- **brakes.sav。**该假设数据文件涉及某生产高性能汽车盘式制动器的工厂的质量控制。该数据文件包含对 8 台专用机床中每一台的 16 个盘式制动器的直径测量。盘式制动器的目标直径为 322 毫米。
- **breakfast.sav。**在一项经典研究中(Green 和 Rao, 1972)，21 名 Wharton School MBA 学生及其配偶被要求按照喜好程度顺序对 15 种早餐食品进行评价，从 1 =他们的喜好根据六种不同的情况加以记录，从“全部喜欢”到“只带饮料的快餐”。
- **breakfast-overall.sav。**该数据文件只包含早餐食品喜好的第一种情况，即“全部喜欢”。
- **broadband_1.sav。**该假设数据文件包含各地区订制了全国宽带服务的客户的数量。该数据文件包含 4 年期间 85 个地区每月的订户数量。
- **broadband_2.sav。**该数据文件和 broadband_1.sav 一样，但包含另外三个月的数据。
- **car_insurance_claims.sav。**在别处被提出和分析的(McCullagh 和 Nelder, 1989)关于汽车损坏赔偿的数据集。平均理赔金额可以当作其具有 gamma 分布来建模，通过使用逆联接函数将因变量的均值与投保者年龄、车辆类型和车龄的线性组合关联。提出理赔的数量可以作为尺度权重。
- **car_sales.sav。**该数据文件包含假设销售估计值、订价以及各种品牌和型号的车辆的物理规格。订价和物理规格可以从 edmunds.com 和制造商处获得。
- **car_sales_uprepared.sav。**这是 car_sales.sav 的修改版本，不包含字段的任何已转换版本。
- **carpet.sav。**在一个常用示例中(Green 和 Wind, 1973)，一家公司非常重视一种新型地毯清洁用品的市场营销，希望检验以下五种因素对消费者偏好的影响—包装设计、品牌名称、价格、优秀家用品 标志和退货保证。包装设计有三个因子水平，每个因子水平因刷体位置而不同；有三个品牌名称（K2R、Glory 和 Bissell）；有三个价格水平；最后两个因素各有两个级别（有或无）。十名消费者对这些因素所定义的 22 个特征进行了排序。变量优选包含对每个特征的平均等级的排序。低排序与高偏好相对应。此变量反映了对每个特征的偏好的总体度量。
- **carpet_prefs.sav。**该数据文件所基于的示例和在 carpet.sav 中所描述的一样，但它还包含从 10 位消费者的每一位中收集到的实际排列顺序。消费者被要求按照从最喜欢到最不喜欢的顺序对 22 个产品特征进行排序。carpet_plan.sav 中定义了变量 PREF1 到 PREF22 包含相关特征的标识符。
- **catalog.sav。**该数据文件包含某编目公司出售的三种产品的假设每月销售数据。同时还包括 5 个可能的预测变量的数据。
- **catalog_seasfac.sav。**除添加了一组从“季节性分解”过程中计算出来的季节性因子和附带的日期变量外，该数据文件和 catalog.sav 是相同的。

- **cellular.sav。**该假设数据文件涉及某便携式电话公司在减少客户流失方面的举措。客户流失倾向分被应用到帐户，分数范围从 0 到 100。得到 50 分或更高分数的帐户可能会更换提供商。
- **ceramics.sav。**该假设数据文件涉及某制造商在确定新型优质合金是否比标准合金具有更高的耐热性方面的举措。每个个案代表对一种合金的单独检验；个案中会记录合金的耐热极限。
- **cereal.sav。**该假设数据文件涉及一份 880 人参于的关于早餐喜好的民意调查，该调查记录了参与者的年龄、性别、婚姻状况以及生活方式是否积极（根据他们是否每周至少做两次运动）。每个个案代表一个单独的调查对象。
- **clothing_defects.sav。**这是关于某服装厂的质量控制过程的假设数据文件。检验员要对工厂中每次大批量生产的服装进行抽样检测并清点不合格的服装的数量。
- **coffee.sav。**这是关于六种冰咖啡的认知品牌形象(Kennedy, Riquier, 和 Sharp, 1996)的数据文件。对于 23 种冰咖啡特征属性中的每种属性，人们选择了由该属性所描述的所有品牌。为保密起见，六种品牌用 AA、BB、CC、DD、EE 和 FF 来表示。
- **contacts.sav。**该假设数据文件涉及一组公司计算机销售代表的联系方式列表。根据这些销售代表所在的公司部门及其公司的秩来对每个联系方式进行分类。同时还记录了最近一次的销售量、最近一次销售距今的时间和所联系公司的规模。
- **creditpromo.sav。**该假设数据文件涉及某百货公司在评价最新信用卡促销的效果方面的举措。为此，随机选择了 500 位持卡人。其中一半收到了宣传关于在接下来的三个月内降低消费利率的广告。另一半收到了标准的季节性广告。
- **customer_dbase.sav。** 该假设数据文件涉及某公司在使用数据仓库中的信息来为最有可能回应的客户提供特惠商品方面的举措。随机选择客户群的子集并为其提供特惠商品，同时记录下他们的回应。
- **customer_information.sav。**该假设数据文件包含客户邮寄信息，如姓名和地址。
- **customer_subset.sav。**来自 customer_dbase.sav 的拥有 80 个个案的子集。
- **customers_model.sav。** 该文件包含某市场营销活动所针对的个人的假设数据。这些数据包括人口统计信息、购物历史摘要和每个人是否响应该活动。每个个案代表单独的个人。
- **customers_new.sav。**该文件包含作为市场营销活动潜在候选人的个人假设数据。这些数据包括人口统计信息和每个人的购物历史摘要。每个个案代表单独的个人。
- **debate.sav。**该假设数据文件涉及在某政治辩论前后对该辩论的参与者所做的调查的成对回答。每个个案对应一个单独的调查对象。
- **debate_aggregate.sav。**该假设数据文件分类汇总了 debate.sav 中的回答。每个个案对应一个辩论前后的偏好的交叉分类。
- **demo.sav。**这是关于购物客户数据库的假设数据文件，用于寄出每月的商品。将记录客户对商品是否有回应以及各种人口统计信息。
- **demo_cs_1.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第一步。每个个案对应不同的城市，并记录地区、省、区和城市标识。
- **demo_cs_2.sav。**该假设数据文件涉及某公司在汇编调查信息数据库方面的举措的第二步。每个个案对应来自第一步中所选城市的不同的家庭单元，并记录地区、省、区、市、子区和单元标识。还包括设计前两个阶段的抽样信息。
- **demo_cs.sav。**该假设数据文件包含用复杂抽样设计收集的调查信息。每个个案对应不同的家庭单元，并记录各种人口统计和抽样信息。

- **dmdata.sav。** 该假设数据文件包含直销公司的人口统计学和购买信息。dmdata2.sav 包含收到测试邮件的联系人子集的信息，dmdata3.sav 包含未收到测试邮件的其余联系人的信息。
- **dietstudy.sav。**该假设数据文件包含对 “Stillman diet” (Rickman, Mitchell, Dingman, 和 Dalen, 1974) 的研究结果。每个个案对应一个单独的主体，并记录其在实行饮食方案前后的体重（磅）以及甘油三酸酯的水平（毫克/100 毫升）。
- **dvdplayer.sav。**这是关于开发新的 DVD 播放器的假设数据文件。营销团队用原型收集了焦点小组数据。每个个案对应一个单独的被调查用户，并记录他们的人口统计信息及其对原型问题的回答。
- **german_credit.sav。**该数据文件取自加州大学欧文分校的 Repository of Machine Learning Databases (Blake 和 Merz, 1998)中的 “German credit” 数据集。
- **grocery_1month.sav。**该假设数据文件是在数据文件 grocery_coupons.sav 的基础上加上了每周购物“累计”，所以每个个案对应一个单独的客户。所以，一些每周更改的变量消失了，而且现在记录的消费金额是为期四周的研究过程中的消费金额之和。
- **grocery_coupons.sav。**该假设数据文件包含由重视顾客购物习惯的杂货连锁店收集的调查数据。对每位顾客调查四周，每个个案对应一个单独的顾客周，并记录有关顾客购物地点和方式的信息（包括那一周里顾客在杂货上的消费金额）。
- **guttman.sav。** Bell (Bell, 1961)创建了一个表，用来阐释可能的社会群体。Guttman (Guttman, 1968)引用了该表的一部分，其中包括五个变量，用于描述以下七个理论社会群体的社会交往、对群体的归属感、成员的物理亲近度以及关系正式性：观众（比如在足球比赛现场的人们）、听众（比如在剧院或听课堂讲座的人们）、公众（比如报纸或电视观众）、组织群体（与观众类似但具有紧密的关系）、初级群体（关系密切）、次级群体（自发组织）及现代社区（因在物理上亲近而导致关系松散并需要专业化服务）。
- **health_funding.sav。**该假设数据文件包含关于保健基金（每 100 人的金额）、发病率（每 10,000 人的比率）以及保健提供商拜访率（每 10,000 的比率）的数据。每个个案代表不同的城市。
- **hivassay.sav。**该假设数据文件涉及某药物实验室在开发用于检测 HIV 感染的快速化验方面的举措。化验结果为八个加深的红色阴影，如果有更深的阴影则表示感染的可能性很大。用 2,000 份血液样本来进行实验室试验，其中一半受到 HIV 感染而另一半没有受到感染。
- **hourlywagedata.sav。**该假设数据文件涉及在政府机关和医院工作的具有不同经验水平的护士的时薪。
- **insurance_claims.sav。**该假设数据文件涉及某保险公司，该公司希望构建一个模型用于标记可疑的、具有潜在欺骗性的理赔。每个个案代表一次单独的理赔。
- **insure.sav。**该假设数据文件涉及某保险公司，该公司正在研究指示客户是否会根据 10 年的人寿保险合同提出理赔的风险因子。数据文件中的每个个案代表一副根据年龄和性别进行匹配的合同，其中一份记录了一次理赔而另一份则没有。
- **judges.sav。**该假设数据文件涉及经过训练的裁判（加上一个体操爱好者）对 300 次体操表演给出的分数。每行代表一次单独的表演；裁判们观看相同的表演。
- **kinship_dat.sav。**Rosenberg 和 Kim (Rosenberg 和 Kim, 1975) 开始分析 15 个亲属关系项（伯母、兄弟、表兄妹、女儿、父亲、孙女、祖父、祖母、孙子、母亲、侄子或外甥、侄女或外甥女、姐妹、儿子和叔叔）。他们让四组大学生（两组女同学，两组男同学）根据相似程度将各项排序。他们让其中的两组同学（一组女同

学，一组男同学）进行了两次排序，第二次排序和第一次排序采取的标准不同。这样，一共得到六组“源”。每个源对应一个 15 × 15 的相似性矩阵，其单元格中的值等于源中的人数减去此源中对象被划分的次数。

- **kinship_ini.sav。**该数据文件包含 kinship_dat.sav 的三维解的初始配置。
- **kinship_var.sav。**该数据文件包含自变量 gender、gener(ation) 和 degree (of separation)，这些变量可用于解释 kinship_dat.sav 的解的维数。具体而言，它们可用来将解的空间限制为这些变量的线性组合。
- **marketvalues.sav。**该数据文件涉及 1999 - 2000 年间 Algonquin，Ill. 地区新的房屋开发中的住房销售。这些销售仅仅来自公众记录。
- **nhis2000_subset.sav。**美国健康访问调查 (NHIS) 是针对美国全体公民的大型人口调查。该调查对美国的具有全国代表性的家庭样本进行了面对面的访问，并获取了每个家庭的成员的健康行为和健康状态的人口统计信息和观察数据。该数据文件包含取自 2000 年调查信息的子集。国家健康统计中心。2000 年美国健康访问调查。公用数据文件和文档。ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/。2003 年发布。
- **ozone.sav。**这些数据包含了用来根据其余变量预测臭氧浓度的六个气象变量的 330 个观察值。在以前的研究人员中，(Breiman 和 Friedman(F), 1985) 和 (Hastie 和 Tibshirani, 1990) 发现了这些变量之间的非线性，这妨碍了标准回归方法。
- **pain_medication.sav。**该假设数据文件包含用于治疗慢性关节炎疼痛的抗炎药的临床试验结果。我们感兴趣的是该药见效的时间以及它和现有药物的比较。
- **patient_los.sav。**该假设数据文件包含被医院确诊为疑似心肌梗塞（即 MI 或“心脏病发作”）的患者的治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **patlos_sample.sav。**该假设数据文件包含在治疗心肌梗塞（即 MI 或“心脏病发作”）期间收到溶解血栓剂的患者的样本治疗记录。每个个案对应一位单独的患者，并记录与其住院期有关的一些变量。
- **polishing.sav。**这是来自 Data and Story Library 的 “Nambeware Polishing Times” 数据文件。该数据文件涉及某金属餐具制造商 (Nambe Mills, Santa Fe, N. M.) 在安排生产计划方面的举措。每个个案代表产品线上的不同项目。并且记录每个项目的直径、抛光时间、价格和产品类型。
- **poll_cs.sav。**该假设数据文件涉及民意测验专家在确定正式立法前公众对法案的支持水平方面的举措。个案对应注册的选民。每个个案记录选民居住的县、镇、区。
- **poll_cs_sample.sav。**该假设数据文件包含在 poll_cs.sav 中列出的选民的样本。该样本是根据 poll.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。请注意，由于该抽样计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (poll_jointprob.sav) 包含联合选择概率。在选取了样本之后，对应于选民人群统计信息及其对提交法案的意见的附加变量将被收集并添加到数据文件。
- **property_assess.sav。**该假设数据文件涉及某县资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应过去一年中县里所出售的资产。数据文件中的每个个案记录资产所在的镇、最后评估资产的评估员、该次评估距今的时间、当时的估价以及资产的出售价格。

- **property_assess_cs.sav。**该假设数据文件涉及某州资产评估员在利用有限的资源不断更新资产价值评估方面的举措。个案对应该州的资产。数据文件中的每个个案记录资产所在的县、镇和区，最后一次评估距今的时间以及当时的估价。
- **property_assess_cs_sample.sav。**该假设数据文件包含在 property_assess_cs.sav 中列出的资产的样本。该样本是根据 property_assess.csplan 中指定的设计来选取的，而且该数据文件记录包含概率和样本权重。在选取了样本之后，附加变量 Current value 将被收集并添加到数据文件。
- **recidivism.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料；如果在第一次被捕后两年内又第二次被捕，则还将记录两次被捕间隔的时间。
- **recidivism_cs_sample.sav。**该假设数据文件涉及某政府执法机构在了解其管辖区域内的屡犯率方面的举措。每个个案对应在 2003 年 6 月期间第一次被捕释放的先前的一名罪犯，并记录其人口统计信息和第一次犯罪的详细资料，及其第二次被捕的数据（如果发生在 2006 年 6 月底之前）。根据 recidivism_cs.csplan 中指定的抽样计划从抽样部门选择罪犯；该计划使用与大小成正比 (PPS) 方法，因此，还有一个文件 (recidivism_cs_jointprob.sav) 包含联合选择概率。
- **rfm_transactions.sav。** 此假设数据文件包含购买交易数据，即每笔交易的购买日期、购买商品和消费金额。
- **salesperformance.sav。**这是关于评估两个新的销售培训课程的假设数据文件。60 名员工被分成 3 组且都接受标准的培训。另外，组 2 接受技术培训；组 3 接受实践教程。在培训课程结束时，对每名员工进行测验并记录他们的分数。数据文件中的每个个案代表一名单独的受训者，并记录其被分配到的组以及测验的分数。
- **satisf.sav。**该假设数据文件涉及某零售公司在 4 个商店位置所进行的满意度调查。总共对 582 位客户进行了调查，每个个案代表一位单独客户的回答。
- **screws.sav。**该数据文件包含关于螺钉、螺栓、螺母和图钉的特征的信息(Hartigan, 1975)。
- **shampoo_ph.sav**这是关于某发制品厂的质量控制的假设数据文件。在规定的时间间隔对六批独立输出的产品进行检测并记录它们的 pH 值。目标范围是 4.5 - 5.5。
- **ships.sav。**在别处被提出和分析的(McCullagh 等。, 1989)关于波浪对货船造成的损坏的数据集。在给定了船的类型、建造工期和服务期后，可以根据以泊松比率发生来为事件计数建模。在因子交叉分类构成的表格中，每个单元格的分类汇总服务月数提供遇到风险的值。
- **site.sav。**该假设数据文件涉及某公司在为扩展业务而选择新址方面的举措。该公司聘请了两名顾问分别对选址进行评估，除了提供长期报告外，他们还要以“前景颇佳”、“前景良好”或“前景不佳”来对每个选址进行总结。
- **smokers.sav。**该数据文件摘自 1998 年全国家庭药物滥用调查并且是美国家庭的概率样本。(http://dx.doi.org/10.3886/ICPSR02934) 因此，分析该数据文件的第一步应该是对数据进行加权以反映总体趋势。
- **stroke_clean.sav。**该假设数据文件包含某医学数据库在经过“数据准备”选项中的过程清理后的状态。
- **stroke_invalid.sav。**该假设数据文件包含某医学数据库的初始状态及一些数据输入错误。

- **stroke_survival。**此假设数据文件涉及正在研究结束缺血性中风后复元计划的患者存活时间的研究人员面临着很多挑战。中风后，记录心肌梗塞、缺血性中风或出血性中风的发生及其时间。样本为左侧截短，因为只包含在中风后管理的复元计划结束后存活的患者。
- **stroke_valid.sav。**该假设数据文件包含在使用“验证数据”过程检查值后，某医学数据库的状态。它仍包含潜在异常个案。
- **survey_sample.sav。**此数据文件包含调查数据，包括人口统计学数据和各种态度测量。它基于 1998 NORC 综合社会调查的变量子集，但某些数据值已经过修改，并添加了其他虚拟变量以供演示用途。
- **telco.sav。**该假设数据文件涉及某电信公司在减少客户群中的客户流失方面的举措。每个个案对应一个单独的客户，并记录各类人口统计和服务用途信息。
- **telco_extra.sav。**该数据文件与 telco.sav 数据文件类似，但删除了 “tenure” 和经对数转换的客户消费变量，代替它们的是标准化的对数转换客户消费变量。
- **telco_missing.sav。** 该数据文件是 telco.sav 数据文件的子集，但某些人口统计数据值已被缺失值替换。
- **testmarket.sav。**该假设数据文件涉及某快餐连锁店为其菜单添加新项目的计划。有三种可能的促销新产品的活动，所以会在多个随机选择的市场中的地点引入新的项目。在每个地点采用不同的促销方式，并记录新项目前四周的每周销售情况。每个个案对应单独地点的一周。
- **testmarket_1month.sav。**该假设数据文件是在数据文件 testmarket.sav 的基础上加上了每周销售“累计”，所以每个个案对应一个单独的地点。所以，一些每周更改的变量消失了，而且现在记录的销售是为期四周的研究过程中的销售之和。
- **tree_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_credit.sav。**该假设数据文件包含人口统计和银行贷款历史数据。
- **tree_missing_data.sav** 该假设数据文件包含具有大量缺失值的人口统计和银行贷款历史数据。
- **tree_score_car.sav。**该假设数据文件包含人口统计和车辆购买价格数据。
- **tree_textdata.sav。**这是一个只有两个变量的样本数据文件，主要打算在指定测量级别和值标签之前显示变量的默认状态。
- **tv-survey.sav。**该假设数据文件涉及由某电视演播室进行的一项关于是否要继续制作一档成功的节目的调查。906 位调查对象被问及他们在各种情况下是否会收看该节目。每行代表一位单独的调查对象；每列代表一种单独的情况。
- **ulcer_recurrence.sav。**此文件包含某项研究的部分信息，该研究旨在比较两种用来防止溃疡复发的治疗的功效。它提供了区间数据的优秀示例并且已在别处被提出和分析(Collett, 2003)。
- **ulcer_recurrence_recoded.sav。**该文件重新组织 ulcer_recurrence.sav 中的信息以允许为研究的每个区间的事件概率建模而不是简单地为研究结束事件概率建模。它已在别处被提出和分析(Collett 等。, 2003)。
- **verd1985.sav。**该数据文件涉及某项调查(Verdegaal, 1985)。该调查记录了 15 个主体对 8 个变量的响应。需要处理的变量被分成 3 个集。数据集 1 包含 年龄 和 婚姻；数据集 2 包含 宠物 和 新闻；数据集 3 包含 音乐 和 居住。宠物被尺度化为多名义而年龄被尺度化为有序；所有其他变量都被尺度化为单名义。

- **virus.sav。**该假设数据文件涉及某因特网服务提供商 (ISP) 在确定病毒对其网络的影响方面的举措。他们从发现病毒到威胁得以遏制这段时间内跟踪其网络上受感染的电子邮件的流量的（近似）百分比。
- **wheeze_steubenville.sav。**这是关于空气污染对儿童健康影响的纵向研究的一个子集 (Ware, Dockery, Spiro III, Speizer, 和 Ferris Jr., 1984)。这些数据包含儿童的气喘状况的重复二分类测量（这些儿童来自 Steubenville，Ohio，年龄为 7 到 10 岁），以及母亲在研究的第一年中是否为吸烟者的固定记录。
- **workprog.sav。**该假设数据文件涉及一份尝试为弱势群体提供较好的工作的政府工作计划。文件后还有一个潜在计划参与者的样本，其中一些参与者是被随机选择来参加该计划的，而其他参与者则不是。每个个案代表一位单独的计划参与者。

附录
B

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided “AS IS”, without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shmtl.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, http://www.winwrap.com.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

参考书目

Bell, E. H. 1961. Social foundations of human behavior:Introduction to the study of sociology. New York: Harper & Row.

Blake, C. L., 和 C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., 和 J. H. Friedman(F). 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, .

Cochran, W. G. 1977. Sampling Techniques, 3rd ed. New York: John Wiley and Sons.

Collett, D. 2003. Modelling survival data in medical research, 2 ed. Boca Raton: Chapman & Hall/CRC.

Cox, D. R., 和 E. J. Snell. 1989. The Analysis of Binary Data, 2nd ed. London: Chapman and Hall.

Green, P. E., 和 V. Rao. 1972. Applied multidimensional scaling. Hinsdale, Ill.: Dryden Press.

Green, P. E., 和 Y. Wind. 1973. Multiattribute decisions in marketing:A measurement approach. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. Psychometrika, 33, .

Hartigan, J. A. 1975. Clustering algorithms. New York: John Wiley and Sons.

Hastie, T., 和 R. Tibshirani. 1990. Generalized additive models. London: Chapman and Hall.

Kennedy, R., C. Riquier, 和 B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. Journal of Targeting, Measurement, and Analysis for Marketing, 5, .

Kish, L. 1965. Survey Sampling. New York: John Wiley and Sons.

Kish, L. 1987. Statistical Design for Research. New York: John Wiley and Sons.

McCullagh, P., 和 J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. 输入: Frontiers in Economics, P. Zarembka，ed. New York: Academic Press.

Murthy, M. N. 1967. Sampling Theory and Methods. Calcutta, India: Statistical Publishing Society.

Nagelkerke, N. J. D. 1991. A note on the general definition of the coefficient of determination. Biometrika, 78:3, .

Price, R. H., 和 D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. Journal of Personality and Social Psychology, 30, .

Rickman, R., N. Mitchell, J. Dingman, 和 J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. Journal of the American Medical Association, 228, .

Rosenberg, S., 和 M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. Multivariate Behavioral Research, 10, .

Särndal, C., B. Swensson, 和 J. Wretman. 1992. Model Assisted Survey Sampling. New York: Springer-Verlag.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, 和 H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents:A longitudinal perspective. British Journal of Psychiatry, 170, .

Verdegaal, R. 1985. Meer sets analyse voor kwalitatieve gegevens (in Dutch). Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, 和 B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. American Review of Respiratory Diseases, 129, .

索引