



南京大學
NANJING UNIVERSITY

第5章 大型研究项目的数据管理

大型研究项目的特点

- 涉及的变量非常多
- 数据的采集不是一次进行，而是分批录入
- 涉及的人员和单位众多
- 项目时间长，核心人员存在流动性

学习目标

- 了解什么是数据字典
- 能熟练应用“定义变量属性”功能
- 能熟练应用“复制数据属性”功能
- 能定义“验证规则”并进行数据验证
- 能标识重复个案
- 能进行双录核查
- 能标识异常个案

主要内容

- 5.1 数据字典
- 5.2 数据核查
- 5.3 数据准备



南京大學
NANJING UNIVERSITY

5.1 数据字典

数据字典简介

- 在大型或者连续性的数据分析项目中，为了保证工作质量，数据处理人员往往会事先定义好一个非常详细的数据格式，包括变量名称、变量类型、变量名标签、值标签、缺失值定义等等，这被称为数据字典，它是使用者定义具体数据文件格式的标准模板。
- 对于正式的数据管理项目而言，数据字典是必备的工具。
- SPSS中使用的数据字典实际上就是预先设定好各种变量格式的空数据文件或有预实验数据的实际数据文件。

数据字典简介

- 与数据字典有关的三个主要功能
 - 定义变量属性
 - 对数据集中已存在的变量进一步定义其属性。
 - 复制数据属性
 - 将定义好的数据字典直接应用到当前文件中。
 - 新建自定义属性
 - 用户自行设定一些特殊的变量属性。

定义变量属性（Define Variable Properties）

- 实际上，该向导的绝大多数功能都可以在变量视图中实现。
- 对于复杂的数据管理项目而言，它的可视化能力可以大大提高工作效率；其次，对初学者而言，使用该向导进行变量的设置也是非常好的选择。
- 进入实现过程
 - “数据” → “定义变量属性”

定义变量属性 (Define Variable Properties)

定义变量属性



扫描数据后，使用此工具来标注变量值并设置其他属性。

选择要扫描的变量。这些变量应为分类变量（标称或有序）以获得最佳结果。您可以在下一个面板中更改测量级别设置。

变量(V):

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- S3. 年龄 [s3]
- S4. 学历 [s4]
- S5. 职业 [s5]
- S7. 婚姻状况 [s7]
- S9. 家庭月收入 [s9]
- C0. 请问您的家庭...
- C0. 请问您的家庭...

要扫描的变量(S):



☐ 将要扫描的个案的数量限定为(L):

☒ 将要显示的值的数量限定为(N):

200

继续

取消

帮助

定义变量属性 (Define Variable Properties)

定义变量属性

×

已扫描的变量列表(C):

无...	测量	角色	变量
<input checked="" type="checkbox"/>			time
<input checked="" type="checkbox"/>			id
<input type="checkbox"/>			s0
<input type="checkbox"/>			s2
<input checked="" type="checkbox"/>			s3
<input type="checkbox"/>			s4
<input type="checkbox"/>			s5
<input type="checkbox"/>			s7
<input type="checkbox"/>			s9
<input type="checkbox"/>			c0_1
<input type="checkbox"/>			c0_2
<input type="checkbox"/>			c0_3
<input type="checkbox"/>			O1
<input type="checkbox"/>			a3
<input type="checkbox"/>			a3a_1
<input type="checkbox"/>			a3a_2
<input type="checkbox"/>			a4
<input type="checkbox"/>			a8
<input type="checkbox"/>			a9

当前变量: time

测量级别(M): 有序(...)

角色(E): 输入

未标记的值: 4

标签(L): 月份

建议(S)

类型(T): 数值

宽度(W): 6

小数(D): 0

属性(B)...

值标签网格(V): 在网格中输入或编辑标签。您可以在底部输入其他值。

	已更改	缺失	计数	值	标签
1	<input type="checkbox"/>	<input type="checkbox"/>	300	200704	
2	<input type="checkbox"/>	<input type="checkbox"/>	304	200712	
3	<input type="checkbox"/>	<input type="checkbox"/>	304	200812	
4	<input type="checkbox"/>	<input type="checkbox"/>	239	200912	
5	<input type="checkbox"/>	<input type="checkbox"/>			

已扫描的个案: 1147

值列表限制: 无

复制属性

到其他变量(F)...

到其他变量(O)...

未标记的值

自动标签(A)

确定

粘贴(P)

重置(R)


取消

帮助

复制数据属性（Copy Data Properties）


- 用于将定义好的数据字典直接应用到当前文件中。
- 在操作时还可以进行自定义，只选择某些变量，或者某些属性进行拷贝，这能大大提高连续性项目对原有资源的利用程度。
- 对于一些特殊的文件属性，如多选题变量集、普通变量集、权重变量设定等，使用该向导进行复制更是会减少许多重复工作。
- 进入实现过程
 - 打开一个已有数据文件，或新建一个数据文件
 - “数据” → “复制数据属性”

复制数据属性 (Copy Data Properties)

 复制数据属性 - 第 1 步 (共 5 步) ×

欢迎使用“复制数据属性向导”。

复制数据属性可将所选变量和数据集属性从打开的数据集或外部 SPSS Statistics 数据文件复制到活动的数据集。

 您还可以将属性从活动数据集中的一个变量复制到另一个变量。

复制的是数据属性，而不是数据值。

选择属性源

☒ 打开的数据集(D)

CCSS_Sample.sav [数据集1]

未标题3 [数据集3]

☐ 外部 SPSS Statistics 数据文件

浏览(B)...

☐ 活动数据集(I) (未标题2 [数据集2])

< 上一步(B)

下一步 >(N)

完成

取消

帮助

复制数据属性（Copy Data Properties）

复制数据属性 - 第 2 步 (共 5 步)



复制数据属性 - 选择源和目标变量

☒ 将所选源数据集变量的属性应用于匹配的活动数据集变量(A)。

☒ 如果尚不存在匹配变量，则在活动数据集中创建匹配变量(C)。

☐ 选择源列表中将复制属性的一个变量以及目标列表中将应用属性的一个或多个变量(R)。

☐ 仅应用数据集属性 - 未选定变量(O)



变量匹配的条件是名称和基本类型（数字或字符串以及字符串长度）相同。将在以下面板上指定要应用的特定属性。右键单击变量以查看其属性。

在属性将被复制到匹配变量的源列表中选择变量，或应在活动数据集中创建，如果它们尚未存在。

源数据集变量(S):

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- S3. 年龄 [s3]
- S4. 学历 [s4]
- S5. 职业 [s5]
- S7. 婚姻状况 [s7]
- S9. 家庭月收入 [s9]
- C0. 请问您的家庭...

选择的变量:38

匹配活动数据集

- time
- id
- s0

匹配变量: 3

要创建的变量: 35

不勾选：选择同名同类型同长度变量（匹配变量）的属性进行复制。

勾选：除了可以将匹配变量的属性复制过来，还可以创建与源数据集匹配的变量。

< 上一步(B)

下一步 >(N)

完成

取消

帮助

复制数据属性 (Copy Data Properties)

复制数据属性 - 第 2 步 (共 5 步)




复制数据属性 - 选择源和目标变量

☐ 将所选源数据集变量的属性应用于匹配的活动数据集变量(A)。

☒ 如果尚不存在匹配变量，则在活动数据集中创建匹配变量(C)。

☒ 选择源列表中将复制属性的一个变量以及目标列表中将应用属性的一个或多个变量(R)。

☐ 仅应用数据集属性 - 未选定变量(O)

 变量匹配的条件是名称和基本类型（数字或字符串以及字符串长度）相同。将在以下面板上指定要应用的特定属性。右键单击变量以查看其属性。

在属性将被复制的源列表选择一个属性，并将属性应用到目标列表中的一个或多个变量。

源数据集变量(S):

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- S3. 年龄 [s3]
- S4. 学历 [s4]
- S5. 职业 [s5]
- S7. 婚姻状况 [s7]
- S9. 家庭月收入 [s9]
- C0. 请问您的家庭...

选择的变量:1

活动数据集变量(D):

- time
- id
- s0

选择的变量:3

将源数据集中一个变量的属性复制到目标数据集的一个或多个变量上。

< 上一步(B)

下一步 >(N)

完成

取消

帮助

复制数据属性 (Copy Data Properties)

复制数据属性 - 第 2 步 (共 5 步)



复制数据属性 - 选择源和目标变量

- ☐ 将所选源数据集变量的属性应用于匹配的活动数据集变量(A)。
 - ☒ 如果尚不存在匹配变量，则在活动数据集中创建匹配变量(C)。
- ☐ 选择源列表中将复制属性的一个变量以及目标列表中将应用属性的一个或多个变量(R)。
- ☐ 仅应用数据集属性 - 未选定变量(O)

变量匹配的条件是名称和基本类型（数字或字符串以及字符串长度）相同。将在以下面板上指定要应用的特定属性。右键单击变量以查看其属性。

在属性将被复制的源列表选择一个属性，并将属性应用到目标列表中的一个或多个变量。

源数据集变量(S):

选择的变量:1

活动数据集变量(D):

选择的变量:3

仅复制文件属性：多
选题变量集、权重设
定等。

< 上一步(B)

下一步 >(N)

完成

取消

帮助

复制数据属性（Copy Data Properties）

- 案例

- 源数据文件：CCSS_Sample.sav, 当前数据文件：copydataproperties.sav (自己创建的，含id、time和s0三个变量，名称类型长度和源数据文件中一致，其他属性默认)
- 要求：进行“复制数据属性”操作

- 实现过程

- “数据” → “复制数据属性”
- “复制数据属性”向导第1步：将CCSS_Sample.sav选择为属性源
- “复制数据属性”向导第2步：“复制数据属性—选择源和目标变量”部分按上几页ppt进行不同的选择

复制数据属性 (Copy Data Properties)

复制数据属性 - 第 2 步 (共 5 步)

×


复制数据属性 - 选择源和目标变量

☒ 将所选源数据集变量的属性应用于匹配的活动数据集变量(A)。

☐ 如果尚不存在匹配变量，则在活动数据集中创建匹配变量(C)。

☐ 选择源列表中将复制属性的一个变量以及目标列表中将应用属性的一个或多个变量(R)。


☐ 仅应用数据集属性 - 未选定变量(O)





变量匹配的条件是名称和基本类型（数字或字符串以及字符串长度）相同。将在以下面板上指定要应用的特定属性。右键单击变量以查看其属性。

在属性将被复制到活动数据集中匹配变量的源列表中选择变量。

源数据集变量(S):


 月份 [time]


 ID [id]


 S0. 城市 [s0]

选择的变量:3

匹配活动数据集

 time

 id

 s0

匹配变量: 3

要创建的变量: 0

< 上一步(B)


下一步 >(N)

完成

取消

帮助

复制数据属性 (Copy Data Properties)


 复制数据属性 - 第 2 步 (共 5 步) ×

复制数据属性 - 选择源和目标变量

☒ 将所选源数据集变量的属性应用于匹配的活动数据集变量(A)。
☒ 如果尚不存在匹配变量，则在活动数据集中创建匹配变量(C)。


☐ 选择源列表中将复制属性的一个变量以及目标列表中将应用属性的一个或多个变量(R)。

☐ 仅应用数据集属性 - 未选定变量(O)

 变量匹配的条件是名称和基本类型（数字或字符串以及字符串长度）相同。将在以下面板上指定要应用的特定属性。右键单击变量以查看其属性。

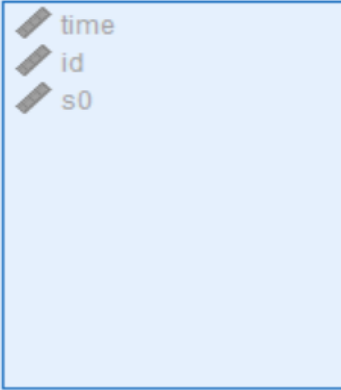
在属性将被复制到匹配变量的源列表中选择变量，或应在活动数据集中创建，如果它们尚未存在。

源数据集变量(S):



选择的变量:38

匹配活动数据集



匹配变量: 3
要创建的变量: 35

< 上一步(B) 下一步 >(N) 完成 取消 帮助

复制数据属性 (Copy Data Properties)

复制数据属性 - 第 2 步 (共 5 步)



复制数据属性 - 选择源和目标变量

☐ 将所选源数据集变量的属性应用于匹配的活动的数据集变量(A)。

☒ 如果尚不存在匹配变量，则在活动数据集中创建匹配变量(C)。

☒ 选择源列表中将复制属性的一个变量以及目标列表中将应用属性的一个或多个变量(R)。

☐ 仅应用数据集属性 - 未选定变量(O)



变量匹配的条件是名称和基本类型（数字或字符串以及字符串长度）相同。将在以下面板上指定要应用的特定属性。右键单击变量以查看其属性。

在属性将被复制的源列表选择一个属性，并将属性应用到目标列表中的一个或多个变量。

源数据集变量(S):

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- S3. 年龄 [s3]
- S4. 学历 [s4]
- S5. 职业 [s5]
- S7. 婚姻状况 [s7]
- S9. 家庭月收入 [s9]
- C0. 请问您的家庭...

选择的变量:1

活动数据集变量(D):

- time
- id
- s0

选择的变量:3

< 上一步(B)

下一步 >(N)

完成

取消

帮助

新建自定义属性

- **新建自定义属性：**在已有11个标准变量属性（名称、类型等）的基础上，用户可根据需要自行设定一些特殊的变量属性。
- **实现过程**
 - 变量视图下，“数据” → “新建定制属性”
 - 设置属性名称
 - 设置默认的属性值
 - 选择相应的变量

新建自定义属性



新建定制属性



为新定制属性选择变量。

变量(V):

- Qa3
- Qa4
- Qa8
- Qa9
- Qa10
- Qa16
- 现状指数 [index1a]
- 预期指数 [index1b]
- time <= 200704 (...)

选择的变量(S):

- 总指数 [index1]



属性名称(N): 基准值

属性值(A): 100

☒ 在数据编辑器中显示属性(D)

显示定义的列表属性(L) >>

确定

取消

帮助



南京大學
NANJING UNIVERSITY

5.2 数据核查

数据核查概述

- 数据核查的基本内容

- 单选题数值核查：数值中不应当出现选项外取值，等
- 开放题数值核查：连续变量应在有效范围内取值，等
- 多选题数值核查：采用多重分类法编码时，同一个选项代码不应当在不同列中重复出现，等
- 逻辑查错：某些变量的值不能同时为9，等

- 数据核查的技术路线

- 任务分解：将各种查错工作归类为若干个基本独立的种类
- 技术实现：对每个分解出的类别给出适当的错误识别规则定义，并采用适当的技术手段实现
- 结果反馈：采用适当的技术手段使得查错结果能够清楚并格式统一地反馈给用户

数据验证模块

- **功能：**实现数据核查
- **实现途径：**用户自定义数据验证规则，也可以通过“数据”→“验证”→“加载预定义规则”加载系统自带的一些验证规则
- **规则分类：**
 - **单变量规则：**一组应用于单个变量的数值检查规则，例如，范围外值的检查，有效值可以表示为一个范围，也可以表示为一个有效值列表
 - **交叉变量规则：**涉及多个变量间逻辑关系的规则，由标记无效值的逻辑表达式定义，可用于单个变量，也可用于变量组合

数据验证模块——案例

- 数据文件：CCSS_bad.sav
- 定义验证规则：（1）性别S2只有1、2两种取值编码；（2）年龄S3取值应当在18-65岁之间；（3）关键题目取值逻辑：A3、A4、A8不应当同时选择9，若都为9应作为废卷处理。
- 实现过程：
 - “数据” → “验证” → “定义规则”
 - “数据” → “验证” → “验证数据”

数据验证模块——案例

定义验证规则

×

单变量规则交叉变量规则

规则(R):

名称	类型
RuleS3	数字
RuleS2	数字

规则定义

名称(M): RuleS2

类型(T): 数字

格式(F): mm/dd/yyyy

有效值(V):

在列表中

值(L):

1

2

☒ 在检查值时忽略大小写(I)

☒ 允许使用用户缺失值(W)

☒ 允许使用系统缺失值(Y)

☒ 允许使用空值(B)

新建(N)

复制(U)

删除(D)

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

定义验证规则

×

单变量规则

交叉变量规则

规则(R):

名称	类型
RuleS3	数字
RuleS2	数字

新建(N)

复制(U)

删除(D)

规则定义

名称(M): RuleS3

类型(T): 数字

格式(F): mm/dd/yyyy

有效值(V):

在范围内

最小(I): 18

最大(X): 65

☒ 允许在范围内使用非整数值(G)

☒ 允许使用用户缺失值(W)

☒ 允许使用系统缺失值(Y)

☒ 允许使用空值(B)

指定最小值、最大值或二者都指定。如果都没指定，所有值将视为在范围内。

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

定义验证规则

×

单变量规则

交叉变量规则

规则(R):

名称

CrossVarRule1

规则定义

名称(M):

CrossVarRule1

逻辑表达式(对于无效个案，计算结果应当为 1) (L):

A3=9 & A4=9 & A8=9

+ - * / ** < > <= >= = ~= & | ~ 0

变量(V):

月份 [time]

ID [id]

S0. 城市 [s0]

S2. 性别 [s2]

S3. 年龄 [s3]

S4. 学历 [s4]

S5. 职业 [s5]

S7. 婚姻状况 [s7]

S9. 家庭月收入 [s9]

C0. 请问您的家庭...

C0. 请问您的家庭...

C0. 请问您的家庭...

插入(S)

函数和特殊变量

函数(F):

Abs

Arsin

Artan

Cos

显示(Y):

全部

插入(I)

描述(C):

新建(N)

复制(P)

删除(D)

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

验证数据

×

变量

基本检查

单变量规则

交叉变量规则

输出

保存

变量(V):

- S0. 城市 [s0]
- S4. 学历 [s4]
- S5. 职业 [s5]
- S7. 婚姻状况 [s7]
- S9. 家庭月收入 [s9]
- C0. 请问您的家庭目前有下列...
- C0. 请问您的家庭目前有下列...
- C0. 请问您的家庭目前有下列...
- O1. 是否拥有家用轿车 [O1]
- A3. 首先, 请问与一年前相比...
- A3a. 您为什么这么说呢? [a3...
- A3a. 您为什么这么说呢? [a3...
- A4. 那么与现在相比, 一年以...
- A8. 那么与现在相比, 您认为...
- A9. 那么您认为一年之后本地...
- A10. 那么与现在相比, 您认为...
- A16. 对于大宗耐用消费品的购...
- 家庭收入2级 [Ts9]
- Qs9
- Qa3
- Qa4

分析变量(A):

- S2. 性别 [s2]
- S3. 年龄 [s3]

个案标识变量(C):

- 月份 [time]
- ID [id]

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

验证数据

×

变量

基本检查

单变量规则

交叉变量规则

输出

保存

分析变量

☒ 标记没有通过下列任一检查的变量(F)

缺失值的最大百分比(M): (适用于所有变量)

单个类别中个案所占的最大百分比(C): (仅适用于分类变量)

计数为 1 的类别的最大百分比(T): (仅适用于分类变量)

最小变异系数(V): (仅适用于刻度变量)

最小标准差(S): (仅适用于刻度变量)

个案标识

☒ 标记不完整的标识(I)

☒ 标记重复的标识(D)

☒ 标记空个案(E) 个案的定义依据(B): 数据集内除标识变量以外的所有变量 ▼

如果所有的相关变量都缺失或为空白，则将个案视为空。

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

验证数据

×

变量

基本检查

单变量规则

交叉变量规则

输出

保存

要向某个变量应用规则，请先选择该变量，然后选中一个或多个规则。

分析变量列表根据数据的扫描显示非缺失值的分布。规则列表显示可以应用到所选变量的所有规则。

分析变量(A):

变量	分布	最小值	最大值	规则
S2. 性别 [s2]		1	3	1
S3. 年龄 [s3]		12	65	1

规则(U):

应用	名称
<input type="checkbox"/>	RuleS3
<input checked="" type="checkbox"/>	RuleS2

显示(D):

所有变量

已扫描的个案数: 300

定义规则(E)...

变量分布

☒ 待扫描个案数限制(L) 个案(C): 5000

重新扫描(S)

 限制扫描的个案数不会影响验证的个案数量。

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

验证数据

×

变量

基本检查

单变量规则

交叉变量规则

输出

保存

要向某个变量应用规则，请先选择该变量，然后选中一个或多个规则。

分析变量列表根据数据的扫描显示非缺失值的分布。规则列表显示可以应用到所选变量的所有规则。

分析变量(A):

变量	分布	最小值	最大值	规则
S2. 性别 [s2]		1	3	1
S3. 年龄 [s3]		12	65	1

规则(U):

应用	名称
<input checked="" type="checkbox"/>	RuleS3
<input type="checkbox"/>	RuleS2

显示(D):

所有变量

已扫描的个案数: 300

定义规则(E)...

变量分布

☒ 待扫描个案数限制(L) 个案(C): 5000

重新扫描(S)

 限制扫描的个案数不会影响验证的个案数里。

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

验证数据

×

变量

基本检查

单变量规则

交叉变量规则

输出

保存

规则(U):

应用	名称	表达式
<input checked="" type="checkbox"/>	CrossVarRule1	A3=9 & A4=9 & A8=9

定义规则(D)...

确定

粘贴(P)

重置(R)

取消

帮助

数据验证模块——案例

Identifier Checks

Duplicate Identifiers

Duplicate Identifiers Group	Number of Duplicates	Cases with Duplicate Identifiers	Identifier	
			月份	ID
1	2	21,137	200704	1

重复标识报告表

数据验证模块——案例

Single-Variable Rules

Rule Descriptions

Rule	Description
RuleS3	Type: Numeric Domain: Range Flag user-missing values: No Flag system-missing values: No Minimum: 18 Maximum: 65 Flag unlabeled values within range: No Flag noninteger values within range: No \$VD.SRule[1]: Rule
RuleS2	Type: Numeric Domain: List Flag user-missing values: No Flag system-missing values: No List: 1, 2 \$VD.SRule[2]: Rule

Rules violated at least once are displayed.

至少违反一次的单变量规则

数据验证模块——案例

Variable Summary

Rule		Number of Violations
S2. 性别	RuleS2	1
	Total	1
S3. 年龄	RuleS3	1
	Total	1

违规变量摘要

数据验证模块——案例

Cross-Variable Rules 至少违反一次的交叉变量规则

Rule	Number of Violations	Rule Expression
CrossVarRule1	1	A3=9 & A4=9 & A8=9

Case Report

Case	Validation Rule Violations		Identifier	
	Single-Variable ^a	Cross-Variable	月份	ID
1	RuleS2 (1)	CrossVarRule1	200704	16
2	RuleS3 (1)		200704	4

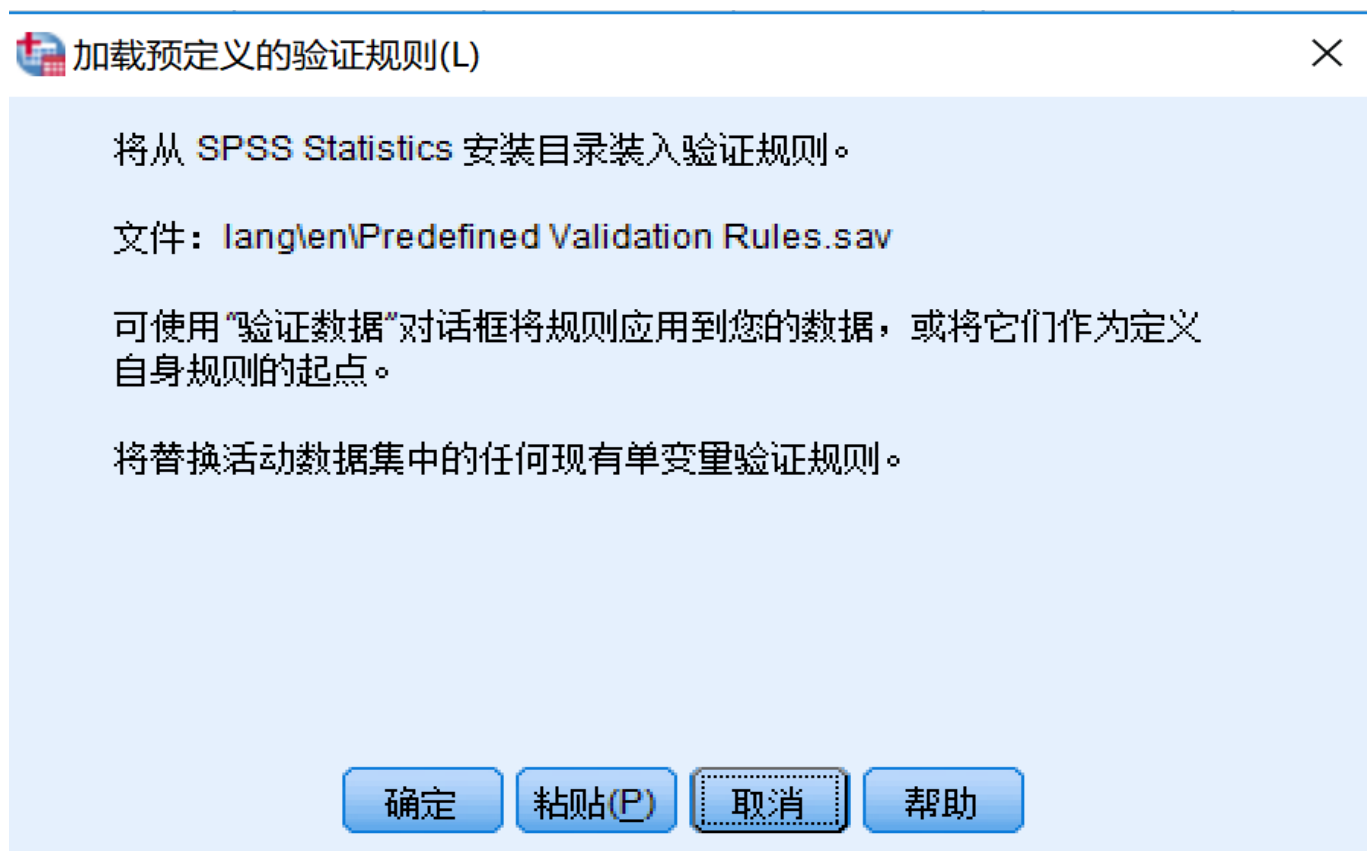
a. The number of variables that violated the rule follows each rule.

违反规则的个案报告

数据验证模块

- 加载预定义规则

- “数据” → “验证” → “加载预定义规则”



标识重复个案（ Identifying Duplicate Cases ）

- **功能：**迅速发现重复的记录，这样的记录可能个别变量值重复，也可能全部变量值重复。
- **进入实现过程**
 - “数据” → “标识重复个案”
- **实现过程**
 - 定义重复个案：一般选择若干个变量，认为个案的这些变量值相同就是重复个案
 - 设置重复记录内部的排序变量和排序形式
 - 勾选“基本个案指示符”：用1标识主个案，用0标识其他的重复个案
 - 勾选“连续计算每个组合中的匹配个案”：为同一个重复记录组内的个案创建序列值为1~n的变量

标识重复个案 (Identifying Duplicate Cases)

标识重复的个案

×

数据文件：
CCSS_bad.sav

- S0. 城市 [s0]
- S2. 性别 [s2]
- S3. 年龄 [s3]
- S4. 学历 [s4]
- S5. 职业 [s5]
- S7. 婚姻状况 [s7]
- S9. 家庭月收入 [s9]
- C0. 请问您的家庭目前...
- C0. 请问您的家庭目前...
- C0. 请问您的家庭目前...
- O1. 是否拥有家用轿车 ...
- A3. 首先, 请问与一年...

定义匹配个案的依据(O):

- 月份 [time]
- ID [id]

在匹配组内的排序标准(O):

排序

☒ 升序(C)

☒ 降序(E)

匹配和分类变量数: 2

要创建的变量

☒ 基本个案指示符 (1=唯一或基本, 0=重复) (I)

☒ 每组中的最后一个个案为基本个案(L)

名称(N): 最后一个基本个案

☐ 每组中的第一个个案为基本个案(H)

☐ 根据指示符的值进行过滤(F)

☐ 连续计算每个组合中的匹配个案
(0=非匹配个案)(S)

名称(M): 匹配顺序

☒ 将匹配个案移至文件顶端(A)

☒ 显示已创建变量的频率(V)

确定

粘贴(P)

重置(R)

取消

帮助

标识重复个案（ Identifying Duplicate Cases ）

Statistics

		所有最后一个 匹配个案的指 示符为主个案	匹配个案的连 续计数
N	Valid	300	300
	Missing	0	0

Frequency Table

所有最后一个匹配个案的指示符为主个案

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	重复个案	1	.3	.3	.3
	主个案	299	99.7	99.7	100.0
	Total	300	100.0	100.0	

双录核查

- **双录核查：**对同一份问卷的两种录入结果进行核查，发现不同之处，从而减少录入错误。
- **进入实现过程**
 - 先打开一个数据集
 - “数据” → “比较数据集”
 - 选择另一个数据集
- **实现过程**
 - 确定要比较的字段
 - 确定个案标识
 - 选择要比较的变量属性
 - 设定不匹配记录的呈现方式

双录核查——案例

- 数据文件：CCSS_bad.sav和CCSS_Sample.sav
- 要求：进行双录核查
- 实现过程
 - 打开文件CCSS_bad.sav
 - “数据” → “比较数据集”，选择文件CCSS_Sample.sav
 - “比较”选项卡中，将“月份”和“ID”选入“个案标识”框，将其余所有变量选入“要比较的字段”框
 - “属性”选项卡，默认设置
 - “输出”选项卡，勾选“将不匹配的个案复制到新数据集”，新数据集名称为“不匹配”

双录核查——案例

 比较数据集 ×

比较 属性 输出

匹配的字段(F):

要比较的字段(C):

 S0. 城市 [s0]

 S2. 性别 [s2]

 S3. 年龄 [s3]

 S4. 学历 [s4]

 S5. 职业 [s5]

个案标识(D):

 月份 [time]

 ID [id]

☒ 排序个案(S)

不匹配的字段(U)...

不匹配(活动) : 0

不匹配(比较) : 1

确定

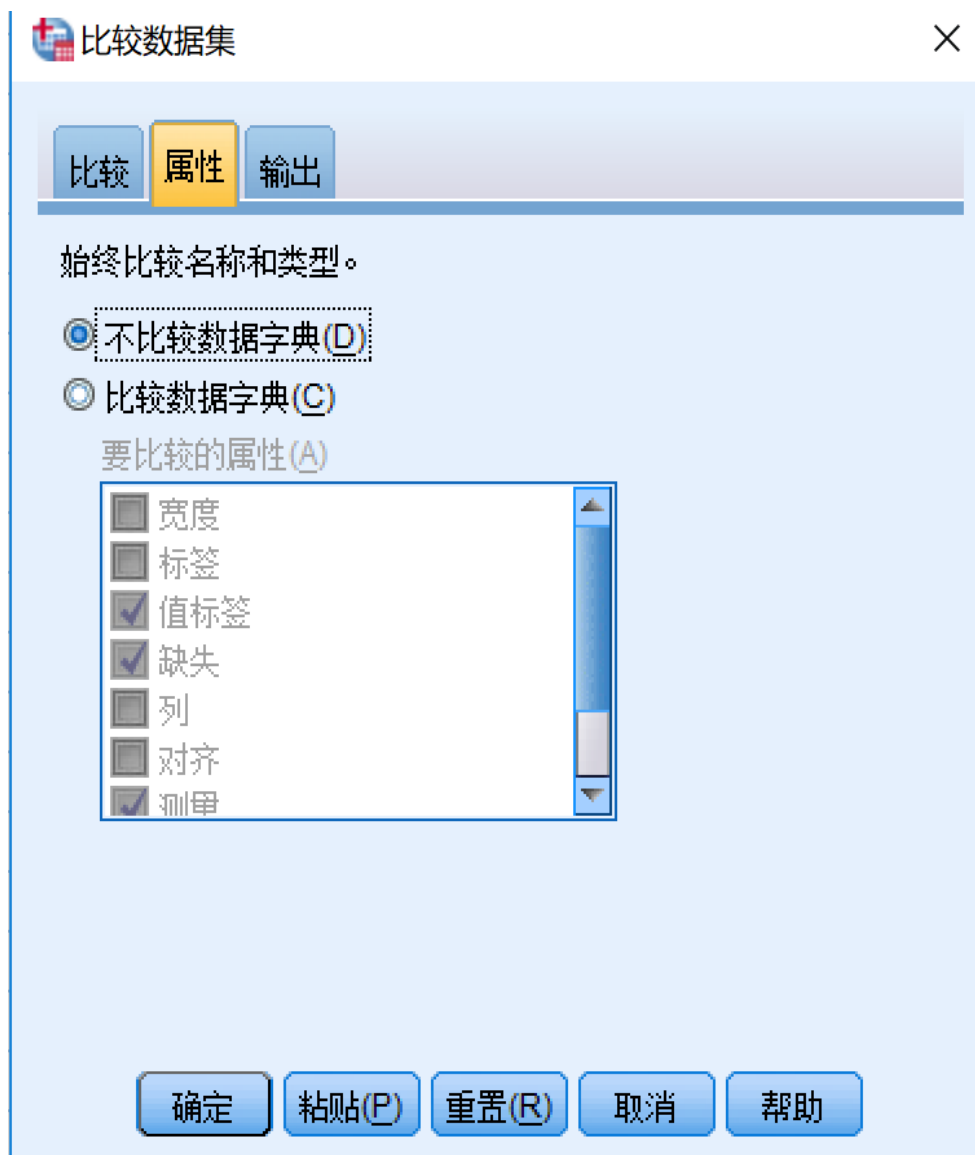
粘贴(P)

重置(R)


取消

帮助

双录核查——案例



双录核查——案例

 比较数据集 ×

比较 属性 **输出**

已保存的变量和数据集

☒ 在新字段中标记不匹配(F)
名称(N) CasesCompare

☐ 将匹配的个案复制到新的数据集(Y)
名称(N)

☒ 将不匹配的个案复制到新的数据集(S)
名称(N) 不匹配

☒ 限制逐项表(L)
报告的最大不匹配数(X): 100

确定 粘贴(P) 重置(R) 取消 帮助

双录核查——案例

Matched Summary

Results Statistics		Datasets	
		Active	Comparison
Cases	Count	300	1147
Cases Compared	Count	299	299
	Percent	99.7%	26.1%
Cases Not Compared	Count	1	848
	Percent	0.3%	73.9%

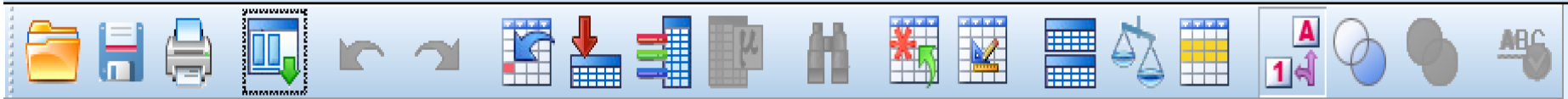
双录核查——案例

Case ID		Row							
time	id	Active	Compare						
200704	1	1	1	(1) 200 (2) 100	(1) 2 (2) 1	(1) 30 (2) 20 (1) 12 (2) 65		(1) 4.00 (2) 3.00	(1) 1 (2) 2
200704	4	5	4						
200704	16	17	16		(1) 3 (2) 1				

(1) is the Active Dataset and (2) is the Comparison Dataset

个案不匹配列表（部分）

文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助(H)

[illegible]



南京大學
NANJING UNIVERSITY

5.3 数据准备

标识异常个案

- 算法原理

- 聚类。将所有个案按照其相似性自动分为若干类。
- 评分。对每一个个案，评估其相对于所属类别的异常度，计算出相应的异常索引。
- 排序。将所有个案按异常索引值降序排列，索引值最高的一部分将被识别为异常个案。
- 报告。对每一个识别出的个案，其相应变量按偏差度指标进行排序。

标识异常个案

- 实现过程

- “数据” → “标识异常个案”
- “变量”选项卡中，选择希望进行异常个案分析的变量以及个案标识变量
- “输出”选项卡中，默认输出异常个案及其异常原因列表
- “保存”选项卡中，可将模型变量保存到数据集
- “缺失值”选项卡中，可控制对用户缺失值和系统缺失值的处理
- “选项”选项卡中，可设定异常个案的标识条件等。

标识异常个案——案例

- 数据文件：CCSS_bad.sav
- 要求：根据index1、index1a和index1b这三个变量识别异常个案
- 实现过程
 - “变量”选项卡中，将index1、index1a和index1b选入“分析变量”框，将ID选入“个案标识变量”框
 - “输出”选项卡中，默认设置
 - “保存”选项卡中，默认设置
 - “缺失值”选项卡中，默认设置
 - “选项”选项卡中，将“异常指标值最高的个案的固定数目”设置为5

标识异常个案——案例

标识异常个案

×

变量

输出

保存

缺失值

选项

变量(V):

- 月份 [time]
- S0. 城市 [s0]
- S2. 性别 [s2]
- S3. 年龄 [s3]
- S4. 学历 [s4]
- S5. 职业 [s5]
- S7. 婚姻状况 [s7]
- S9. 家庭月收入 [s9]
- C0. 请问您的家庭目前有下列还贷支出吗: 房...
- C0. 请问您的家庭目前有下列还贷支出吗: 车...
- C0. 请问您的家庭目前有下列还贷支出吗: 其...
- O1. 是否拥有家用轿车 [O1]
- A3. 首先, 请问与一年前相比, 您的家庭现在...
- A3a. 您为什么这么说呢? [a3a_1]
- A3a. 您为什么这么说呢? [a3a_2]
- A4. 那么与现在相比, 一年以后您的家庭经济...
- A8. 那么与现在相比, 您认为一年以后本地区...

分析变量(A):

- 总指数 [index1]
- 现状指数 [index1a]
- 预期指数 [index1b]



个案标识变量(C):

- ID [id]

要更改变量的测量级别, 在变量列表中右键单击。

确定

粘贴(P)

重置(R)

取消

帮助

标识异常个案——案例



标识异常个案



变量

输出

保存

缺失值

选项

异常个案的标识条件

- ☐ 具有最高异常指标值的个案所占的百分比(E)

百分比(G): 5

- ☒ 具有最高异常指标值的个案的固定数量(E)

数量(B): 5

- ☒ 仅标识异常指标值符合或超过最小值的个案(I)

分界值(T): 2

对等组的数量

最小(N): 1

最大(M): 15

最大的原因数量(X):

1

如果保存了原因 变量，指定要在输出中报告和添加到活动数据集中的原因数量。如果超过分析变量的数量，将值向下调整。

确定

粘贴(P)

重置(R)

取消

帮助

标识异常个案——案例

Anomaly Case Index List

Case	id	Anomaly Index
156	156	6.710
140	140	6.470
258	258	5.832
230	230	5.619
43	42	4.712

标识异常个案——案例

Anomaly Case Peer ID List

Case	id	Peer ID	Peer Size	Peer Size Percent
156	156	4	40	13.3%
140	140	2	37	12.3%
258	258	1	79	26.3%
230	230	1	79	26.3%
43	42	2	37	12.3%

标识异常个案——案例

Anomaly Case Reason List

Reason: 1

Case	id	Reason Variable	Variable Impact	Variable Value	Variable Norm
156	156	index1b	.497	145.29	117.7418
140	140	index1	.416	46.86	88.4506
258	258	index1b	.533	12.11	72.0297
230	230	index1	.427	31.24	75.9316
43	42	index1a	.507	.00	70.1908

THE END