

第3章 变量级别的数据管理

- 数据管理:

- 对数据文件的结构进行重新调整或转换，以便适合于相应的统计方法。

- 数据管理包括的内容:

- 计算新变量、变量值的分组合并、变量的重新编码、变量的编秩，加权个案、文件合并与拆分、分类汇总、数据文件重组、等等。

- 数据管理有两个级别：

- 变量级别（Transform菜单即“转换”菜单）

- 计算变量：对变量进行四则运算生成新变量
 - 变量转换：Recode(重新编码)、Visual Bander(可视离散化)、Count(计数)、Rank Cases(排秩个案)、Automatic Recode(自动重新编码)等

- 文件级别（Data菜单即“数据”菜单）

- 简单命令：插入变量，插入个案，复制数据集，等
 - 常用过程：个案排序、筛选和加权，拆分文件，等
 - 数据重组：长宽型数据相互转换，数据转置
 - 文件合并：将几个数据文件合并为一个数据文件

学习目标

- 能利用 “Compute” ([计算变量](#)) 计算新变量
- 能利用 “Recode” ([重新编码](#)) 对变量值进行分组
- 能利用 “Visual Bander” ([可视离散化](#)) 对连续型变量进行可视化分组
- 能利用 “Automatic Record” ([自动重新编码](#)) 将字符变量转化成数值变量
- 能利用 “Rank Cases” ([个案排秩](#)) 计算秩次

主要内容

- 3.1 变量赋值
- 3.2 已有变量值的分组合并
- 3.3 连续变量的离散化
- 3.4 自动重编码、编秩与数值计数



南京大學
NANJING UNIVERSITY

3.1 变量赋值

“变量赋值”的概念和应用场景

- **变量赋值：**在原有数据的基础上，根据用户的要求，使用SPSS算术表达式及函数，对所有记录或满足SPSS条件表达式的某些记录进行四则运算，并将结果存入一个用户指定的变量中。
- **应用场景：**计算新变量，或者给老变量赋值。

SPSS算术表达式和SPSS函数

SPSS算术表达式：由常量、SPSS变量名、SPSS算术运算符、圆括号等组成的式子，参与运算的数据类型和最终结果均为数值型。

SPSS函数：SPSS为数据处理提供了百余种系统函数，包括：算术函数、统计函数、分布函数、逻辑函数、字符串函数、日期时间函数、缺失值函数、等。

SPSS条件表达式与逻辑表达式

- **SPSS条件表达式：**为了选择感兴趣的记录进行运算而设置的表达式。表达式中常用到的关系运算符有：<、>、<=、>=、=、~=。
- **SPSS逻辑表达式：**具有& (AND)、| (OR)、~ (NOT) 的SPSS条件表达式。

“计算变量” 过程

- “变量赋值” 通过 “计算变量” 过程实现：
 - “转换” → “计算变量”



“计算变量” 案例

- 数据文件：CCSS_Sample.sav
- 要求：将受访对象按年龄段分组：18-34、35-54、55-65，生成新变量TS3存放组号：1、2、3。
 - 说明：这是一个典型的变量值分组问题，一般采用Recode(重新编码)实现，此处尝试通过“计算变量”实现。
- 实现过程：
 - “计算变量” → “目标变量名” 设为TS3，“数字表达式” 设为1 → 确认；
 - “计算变量” → “数字表达式” 设为2 → “如果” 按钮点击进入后设置条件表达式 “s3>=35 & s3<=54” → “继续” → “确认” → “确认”；
 - “计算变量” → “数字表达式” 设为3 → “如果” 按钮点击进入后设置条件表达式 “s3>=55” → “继续” → “确认” → “确认”。

目标变量(T):

TS3

=

数字表达式(E):

1

类型与标签(L)...

- 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. 请问您的家庭目...
- C0. 请问您的家庭目...



+	<	>	7	8	9
-	<=	>=	4	5	6
*	=	~=	1	2	3
/	&		0	.	
**	~	()	删除		



函数组(G):

- 全部
- 算术
- CDF 与非中心 CDF
- 转换
- 当前日期/时间
- 日期运算
- 日期创建

函数和特殊变量(F):

如果(I)... (可选的个案选择条件)

确定

粘贴(P)

重置(R)

取消

帮助

目标变量(T):

TS3

=

数字表达式(E):

2

类型与标签(L)...

- 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. 请问您的家庭目...
- C0. 请问您的家庭目...
- TS3



+	<	>	7	8	9
-	<=	>=	4	5	6
*	=	~=	1	2	3
/	&		0		.
**	~	()	删除		



函数组(G):

- 全部
- 算术
- CDF 与非中心 CDF
- 转换
- 当前日期/时间
- 日期运算
- 日期创建

函数和特殊变量(F):

如果(I)...

s3 >= 35 & s3 <= 54

目标变量(T):

TS3

=

数字表达式(E):

3

类型与标签(L)...



- 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. 请问您的家庭目...
- C0. 请问您的家庭目...
- TS3

+	<	>	7	8	9
-	<=	>=	4	5	6
*	=	~=	1	2	3
/	&		0		.
**	~	()	删除		



函数组(G):

- 全部
- 算术
- CDF 与非中心 CDF
- 转换
- 当前日期/时间
- 日期运算
- 日期创建
- ...

函数和特殊变量(F):

如果(I)...

s3 >= 55

确定

粘贴(P)

重置(R)

取消

帮助



南京大學
NANJING UNIVERSITY

3.2 已有变量值的分组合并

应用场景

- 将连续变量转换为等级变量
 - 将百分制的成绩分为优、良、中、差4个等级；
 - 将年龄分为3组。
- 将分类变量不同的变量等级进行合并
 - 将优、良、中合并为通过，差为不通过。

Recode (重新编码)

- Recode过程：
 - Recode into same variable (重新编码为相同变量)
 - Recode into different variable (“重新编码为不同变量”，较常用)
- 对连续变量进行分组需注意组边界取值
 - 组边界值归为哪个类别？“先下手为强”

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- 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...
- O1. 是否拥有家用...
- A3. 首先, 请问与...
- A3a. 您为什么这...



数字变量 -> 输出变量:

s3 -> TS3

旧值和新值(O)...

如果(I)... (可选的个案选择条件)

输出变量

名称(N):

TS3

标签(L):

更改(H)

旧值

☒ 值(V):

☐ 系统缺失(S)

☐ 系统或用户缺失(U)

☐ 范围(N):

到(T)

☐ 范围，从最低到值(G):

☒ 范围，从值到最高(E):

☐ 所有其他值(O)

新值

☒ 值(L):

☐ 系统缺失(Y)

☐ 复制旧值(P)

旧 → 新(D):

添加(A)

更改(C)

删除(R)

☐ 输出变量为字符串(B) 宽度(W):

8

☐ 将数值字符串移动为数值(M) ('5'→5)

继续

取消

帮助

案例

- 数据文件：CCSS_Sample.sav
- 要求：将受访对象按年龄段分组：18-34、35-54、55-64，生成新变量TS3存放组号：1、2、3。
- 实现过程：
 - 设置输入变量和输出变量：输入变量为s3，输出变量为TS3；（注意要点击“更改”按钮）；
 - 设置旧值和新值的对应关系。

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- 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...
- O1. 是否拥有家用...
- A3. 首先, 请问与...
- A3a. 您为什么这...



数字变量 -> 输出变量:

s3 -> TS3

旧值和新值(O)...

如果(I)... (可选的个案选择条件)

输出变量

名称(N):

TS3

标签(L):

更改(H)

旧值

☐ 值(V):

☐ 系统缺失(S)

☐ 系统或用户缺失(U)

☐ 范围(N):

到(T)

☐ 范围, 从最低到值(G):

☐ 范围, 从值到最高(E):

☒ 所有其他值(O)

新值

☐ 值(L):

☐ 系统缺失(Y)

☒ 复制旧值(P)

旧 -> 新(D):

Lowest thru 34 -> 1

35 thru 54 -> 2

55 thru Highest -> 3

ELSE -> Copy

添加(A)

更改(C)

删除(R)

☐ 输出变量为字符串(B)

宽度(W):

8

☐ 将数值字符串移动为数值(M) ('5' -> 5)

继续

取消

帮助

3.3 连续变量的离散化

应用场景


- 等距分组
- 等样本量分组
- 有一定的可视化要求

“可视分箱化”(Visual Bander)过程

可视分箱

×

已扫描的变量列表(C):

 S3. 年龄 [s3]

名称:

标签:

当前变量:

s3

S3. 年龄

分箱化的变量(B):

S3. 年龄 (已分箱化)

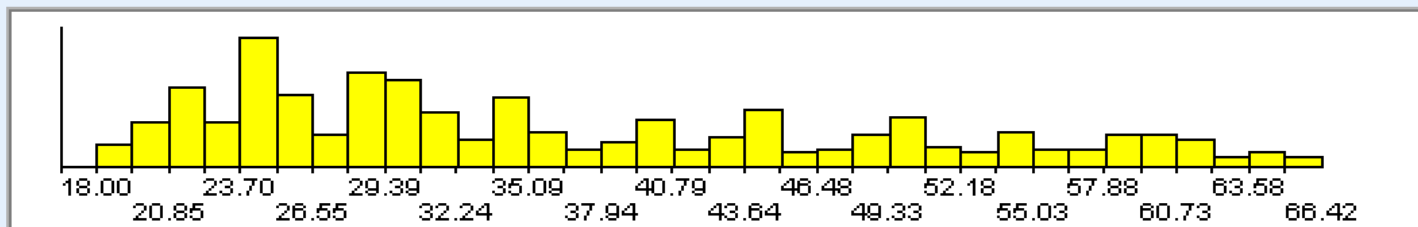
最小:

18

非缺失值

最大:

65



输入间隔分割点或单击“生成分割点”以自动创建区间。例如，值为 10 的分割点所定义的区域是起始于上一个区间之上，结束于 10。

网格(G):

	值	标签
1	HIGH	
2		

上端点

☒ 包含(I) (\leq)

☐ 排除(E) ($<$)

生成分割点(M)...

生成标签(A)

☐ 反向刻度(S)

已扫描个案:

1147

缺失值:

0

复制分箱

从其他变量(F)...

到其他变量(T)...

确定


粘贴(P)

重置(R)

取消

帮助

“可视分箱化”（ Visual Bander ）过程

 生成分割点 ×

☒ 等宽度间隔(E)

间隔 - 至少填充两个字段

第一个分割点的位置(F):

分隔点数量(N):

宽度(W):

最后一个分隔点的位置:

☐ 基于已扫描个案的等百分位(U)

间隔 - 填充任一字段

分隔点数量(N):


宽度(%) (W):

☐ 基于已扫描个案的平均和选定标准差处的分割点(C)

☐ +/- 1 标准差

☐ +/- 2 标准差

☐ +/- 3 标准差

 选择“应用”将会用指定的设置替换当前的分割点定义。

最终间隔将包含所有剩余值: N 个分割点将生成 N+1 个区间。

案例

- 数据文件：CCSS_Sample.sav
- 要求：将s3年龄变量值分为10组，要求等间距。
- 实现过程：
 - 选择要离散的变量；
 - 为新生成的变量命名；
 - 生成分割点；
 - 生成标签。



选择值将分组为分箱的变量。单击继续将扫描数据。

下方所列变量包含所有数值有序变量和刻度变量。

变量(V):

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

要分箱的变量(B):

S3. 年龄 [s3]




☐ 将要扫描的个案的数量限定为(L):

继续

取消

帮助

已扫描的变量列表(C):

 S3. 年龄 [s3]

已扫描个案: 1147

缺失值: 0

复制分箱

从其他变量(F)...

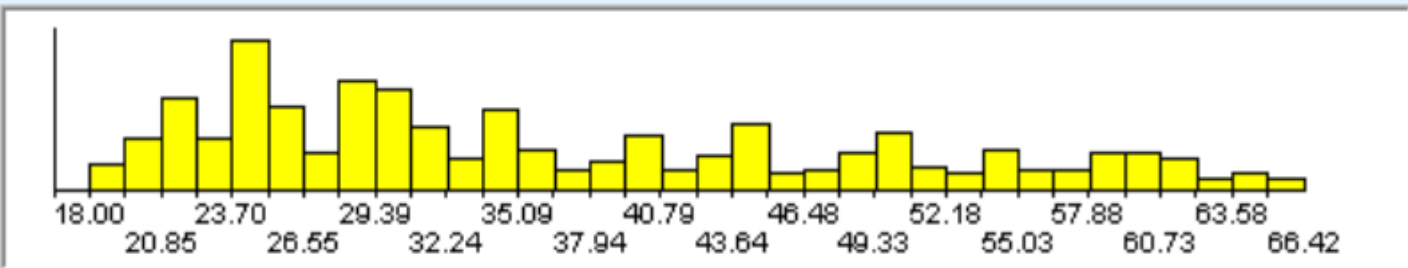
到其他变量(T)...


名称: 标签:

当前变量: s3 S3. 年龄

分箱化的变量(B): s3new S3. 年龄 (已分箱化)

最小: 18 非缺失值 最大: 65



 输入间隔分割点或单击“生成分割点”以自动创建区间。例如，值为 10 的分割点所定义的区间是起始于上一个区间之上，结束于 10。

网格(G):

	值	标签
1	HIGH	
2		

上端点

☐ 包含(I) (<=)

☒ 排除(E) (<)

生成分割点(M)...

生成标签(A)

☐ 反向刻度(S)

☒ 等宽度间隔(E)

间隔 - 至少填充两个字段

第一个分割点的位置(F): 23

分隔点数量(N): 9

宽度(W): 5

最后一个分隔点的位置: 63

☐ 基于已扫描个案的等百分位(U)

间隔 - 填充任一字段

分隔点数量(N):

宽度(%) (W):

☐ 基于已扫描个案的平均和选定标准差处的分割点(C)☐ ± 1 标准差☐ ± 2 标准差☐ ± 3 标准差

选择“应用”将会用指定的设置替换当前的分割点定义。


最终间隔将包含所有剩余值: N 个分割点将生成 N+1 个区间。

应...

取消

帮助

已扫描的变量列表(C):

 S3. 年龄 [s3]

已扫描个案: 1147

缺失值: 0

复制分箱

从其他变量(F)...

到其他变量(T)...

名称:

当前变量: s3

分箱化的变量(B): s3new

最小: 18

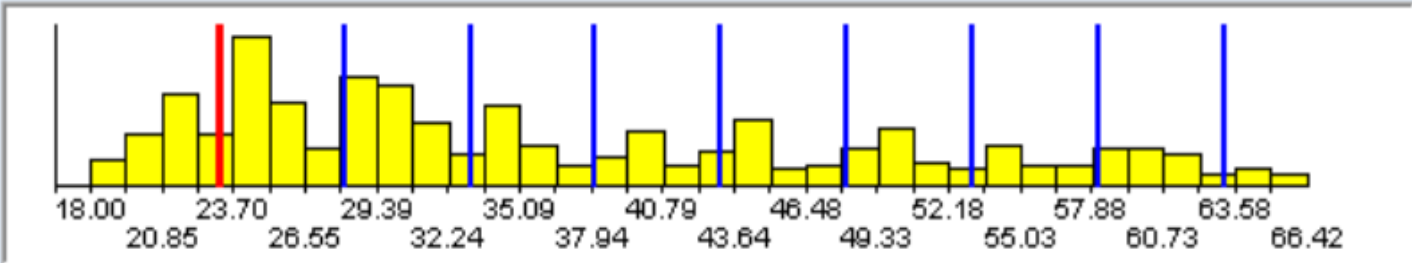
非缺失值


最大: 65

标签:

S3. 年龄

S3. 年龄 (已分箱化)



 输入间隔分割点或单击“生成分割点”以自动创建区间。例如，值为 10 的分割点所定义的区间是起始于上一个区间之上，结束于 10。

网格(G):

	值	标签
1	23.0	< 23
2	28.0	23 - 27
3	33.0	28 - 32
4	38.0	33 - 37
5	43.0	38 - 42
6	48.0	43 - 47
7	53.0	48 - 52
8	58.0	53 - 57

上端点

☐ 包含(I) (<=)

☒ 排除(E) (<)

生成分割点(M)...

生成标签(A)

☐ 反向刻度(S)

“最优分箱化” 过程

- 应用场景

- 因变量为分类变量，而自变量为连续变量，此时需要对自变量进行“最优化分箱”。
- 例如，想要分析“学历”和“年龄”的关系时，可用此过程。

- 实现过程

- 选择要离散的变量和相应的分类变量；
- “输出”选项卡全部勾选；
- “保存”选项卡勾选第1项。



变量

输出

保存

缺失值

选项

变量(V):

- 月份 [time]
- ID [id]
- S0. 城市 [s0]
- S2. 性别 [s2]
- 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...
- C0. 请问您的家庭目前有下列还...
- C0. 请问您的家庭目前有下列还...
- O1. 是否拥有家用轿车 [O1]
- A3. 首先, 请问与一年前相比, 您...
- 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...
- A8. 那么与现在相比, 您认为一年...
- A9. 那么您认为一年之后本地区的...



要分箱的变量(B):



S3. 年龄 [s3]



根据以下项优化分箱(O):



S4. 学历 [s4]



为存储和标称优化变量选择一个或多个刻度变量。

已选择的存储使存储变量和优化变量之间的联系最大化。

可以在“保存”选项卡上保存包括已分箱化的数据值和/或分箱化规则的变量。

确定

粘贴(P)

重置(R)

取消

帮助

Binning Summary

S3. 年龄

Bin	End Point		Number of Cases by Level of S4. 学历					
	Lower	Upper	初中/技校或以下	高中/中专	大专	本科	硕士或以上	Total
1	^a	40	56	149	236	233	48	722
2	40	^a	98	164	95	59	9	425
Total			154	313	331	292	57	1147

Each bin is computed as Lower \leq S3. 年龄 $<$ Upper.

a. Unbounded



南京大學
NANJING UNIVERSITY

3.4 自动重编码、编秩与数值计数

“自动重编码”的应用场景

- 自动将字符变量转换为数值变量
- 自动将数值变量重编码
- 实现过程中将按原变量值的字母排序或者大小生成新变量，而新变量的值就是原值的次序

“自动重编码” (Automatic Code) 过程

- 数据文件: CCSS_Sample.sav
- 要求: 将变量s0自动重新编码
- 结果:

原值	新值
100	1
200	2
300	3



自动重新编码



- 月份 [time]
- ID [id]
- 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...
- C0. 请问您的家庭目...
- TS3



变量 -> 新名称(V)

s0-->s0new

新名称(N): s0new

添加新名称(A)

重新编码的起点

☒ 最低值(L) ☐ 最高值(H)

☐ 对所有变量使用相同的重新编码方案(U)

☐ 将空字符串值视为用户缺失值(B)

模板

☐ 从文件应用模板(T): 文件(F)...

☐ 将模板另存为(S): 文件(I)...

确定

粘贴(P)

重置(R)

取消

帮助

```
AUTORECODE VARIABLES=s0
```

```
/INTO s0new
```

```
/PRINT.
```

```
s0 into s0new (S0. 城市)
```

```
Old Value  New Value  Value Label
```

```
100        1 100北京
```

```
200        2 200上海
```

```
300        3 300广州
```

“个案等级排序” (Rank Cases) 过程

- 应用场景：给变量值排序，给出秩次号
- 结点：相同的观测值形成一个结点
- 在结点处秩次的处理方法：
 - 都取最小秩次
 - 都取最大秩次
 - 都取平均秩次
 - 当做一个记录处理

“个案等级排序” (Rank Cases) 过程

- 数据文件: CCSS_Sample.sav
- 要求: 根据s2性别分组计算s3年龄的秩次
- 实现过程:
 - 设置需计算秩次的变量和分组变量;
 - 选择将秩1分配给最大值还是最小值;
 - 设置结点的处理方法。

月份 [time]
ID [id]
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...
C0. 请问你的家庭



变量(V):

S3. 年龄 [s3]

排序标准(B):

S2. 性别 [s2]

等级的类型(K)...

结(T)...

将等级 1 指定给

☐ 最小值(S)

☒ 最大值(L)

☒ 显示摘要表(D)

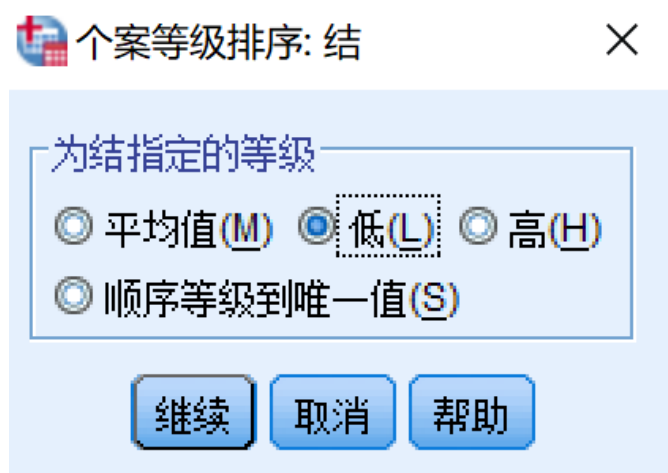
确定

粘贴(P)

重置(R)

取消

帮助



Created Variables^a

Source Variable	Function	New Variable	Label
s3 ^b	Rank	Rs3	Rank of s3 by s2

a. Lowest rank of tied values is used for ties.

b. Ranks are in descending order.

“对个案内的值计数”（Count）过程

- 应用场景：统计每个个案内满足条件的变量数
 - 如计算每位学生的优秀科目数
- 数据文件：CCSS_Sample.sav
- 要求：标识出 $s3 \geq 55$ 的个案
- 实现过程：
 - 设置目标变量；
 - 选择需要计数的数值变量；
 - 设置计数要求。



计算个案内值的出现次数



目标变量(T):

s355

目标标签(L):

数字变量:



S3. 年龄 [s3]



月份 [time]



ID [id]



S0. 城市 [s0]



S2. 性别 [s2]



S4. 学历 [s4]



S5. 职业 [s5]



S7. 婚姻状况 [s7]



S9. 家庭月收入 [s9]



C0. 请问你的家庭



定义值(D)...

如果(I)...

(可选的个案选择条件)

确定

粘贴(P)

重置(R)

取消

帮助



值

☒ 值(V):☐ 系统缺失(S)☐ 系统或用户缺失(U)☐ 范围(N):

到(T):

☐ 范围，从最低到值(G):☐ 范围，从值到最高(E):

要统计的值(O):

55 thru Highest

添加(A)

更改(C)

删除(R)

继续

取消

帮助

练习

- 自行完成本章中涉及的对CCSS_Sample.sav的数据管理操作
- 针对SPSS自带数据Employee data.sav, 进行以下练习:
 - 根据变量bdata生成一个新变量“年龄”（提示：可使用函数XDATE.YEAR()）
 - 根据jobcat分组计算salary的秩次
 - 生成新变量grade, 当salary<20000时取值为d, 在20000~50000范围内时取值为c, 在50000~100000范围内时取值为b, 大于等于100000时取值为a

THE END