SPSSAU 数据科学分析平台

生成变量

目录

数据清理时经常会使用到生成变量功能，SPSSAU 提供的生成变量功能，共包括常用、量纲处理、科学计算、汇总处理、日期相关处理和其它共计六类，操作上：

- ✓ 选中‘标题’，可配合 **ctrl** 或者 **shift** 键同时选择多个标题，批量处理；
- ✓ 只需要选中‘标题’，右侧进行设置然后‘确认处理’即可。
- ✓ 生成变量功能截图和功能汇总如下：

项	说明
常用	包括 8 项，分别是：平均值、求和、虚拟(哑)变量、*z* 标准化、中心化、乘积（交互项）、自然对数（Ln）和 10 为底对数
量纲处理	包括 15 项，分别是：归一化、均值化、正向化、逆向化、适度化、区间化、初值化、最小值化、最大值化、求和归一化、平方和归一化、固定值化、偏固定值化、近区间化、偏区间化
科学计算	包括 7 项，分别是平方、根号、绝对值、倒数、相反数、三次方和取整
汇总处理	包括 4 项，分别是最大值、最小值、中位数和计数
日期相关处理	包括 5 项，分别是日期处理、日期相减、滞后处理、差分处理和季节差分

其它	包括 10 项，分别是样本编号、Box-Cox 变换、秩、缩尾处理、截尾处理、Johnson 转换、排名、相除、相减、非负平移

常用

X1	X2	平均值	求和	乘积（交互项）
1	2	1.5	3	2
2	3	2.5	5	6
3	4	3.5	7	12
3	5	4.0	8	15

比如数据中有 X1 和 X2，此时求该两项的平均值或求和，分别为第 3 和第 4 列。如果是乘积（分析上通常称其为交互项），即选中项之间进行相乘，见表格中第 5 列。除此之外，常用功能中还包括 z 标准化、中心化、自然对数和 10 为底对数，如下表格：

X1	z 标准化	中心化	自然对数	10 为底对数
1	-1.3056	-1.2500	0	0
2	-0.2611	-0.2500	0.6931	0.3010
3	0.7833	0.7500	1.0986	0.4771
3	0.7833	0.7500	1.0986	0.4771

将 X1 进行 z 标准化，进行 z 标准化处理后，数据变成平均值为 0，标准差为 1，关于 z 标准化，其公式如下。

$$z标准化=\frac{x-\overline{x}}{Std},\overline{x}表示平均值,Std表示标准差$$

中心化处理后，数据变成平均值为 0，其公式如下。

$$中心化=x-\overline{x},\overline{x}表示平均值$$

通常情况下，如果数字非常大，可使用自然对数或 10 为底对数，将数据进行压缩。

量纲处理

量纲处理指将数据压缩在一定范围内，但在部分量纲处理方式时，其还可以对数据的方向进行统一（比如违约率数字越小越好，使用逆向化处理将其转换成数字越大越好），量纲处理是一系列处理的统称，SPSSAU 共包括 15 项，如下：

项	说明	注意
归一化	处理后数字介于[0,1]之间	无
均值化	处理后数字大小表示平均值的倍数	通常仅针对大于 0 的数据

正向化	处理后数字介于[0,1]之间，并且数字越大越好	无
逆向化	处理后数字介于[0,1]之间，并且数字越大越好	无
适度化	处理后越接近某个数字越好	无
区间化	处理后数字介于设定的固定区间内	无
初值化	处理后数字大小代表初值(第 1 个数字)的倍数	通常仅针对大于 0 的数据
最小值化	处理后数字大小代表最小值的倍数	通常仅针对大于 0 的数据
最大值化	处理后数字大小代表最大值的倍数	通常仅针对大于 0 的数据
求和归一化	处理后数字大小代表占的比例	通常仅针对大于 0 的数据
平方和归一化	处理后数字大小代表‘相对占比’	通常仅针对大于 0 的数据
固定值化	越接近某个数字越好，处理后数字介于[0,1]之间	无
偏固定值化	越远离某个数字越好，处理后数字介于[0,1]之间	无
近区间化	越接近某个区间越好，处理后数字介于[0,1]之间	无
偏区间化	越远离某个区间越好，处理后数字介于[0,1]之间	无

其具体公式如下：

归一化时，分子为 X 值与最小值的差值，分母是最大值与最小值的差值其为一个固定值。分子取最大时其与分母相等，此时归一化值为 1，分子的最小值为 0，此时归一化值为 0，因而归一化处理后，数字被压缩在[0,1]之间。

$$\text{归一化:}\frac{x-x_{\text{Min}}}{x_{\text{Max}}-x_{\text{Min}}}$$

均值化时，分子是 X 值，分母是 X 值的平均值，其意义为平均值的倍数，均值化通常仅针对全部均大于 0 的数字。

$$\text{均值化:}\frac{x}{\overline{x}},\overline{x}\text{表示平均值}$$

正向化时，分子为 X 值与最小值的差值，分母是最大值与最小值的差值其为一个固定值。分子取最大时其与分母相等，此时归一化值为 1，分子的最小值为 0，此时归一化值为 0，因而归一化处理后，数据被压缩在[0,1]之间，正向化与归一化的公式完全一致，其实际意义为将数字进行压缩在[0,1]之间，并且保持其数字的相对大小意义不变化。

$$\text{正向化:}\frac{x-x_{\text{Min}}}{x_{\text{Max}}-x_{\text{Min}}}$$

正向化时，分子为最大值与 X 值的差值，分母是最大值与最小值的差值，其为一个固定值。X 取最大值时，此时分子最大即归一化值最大，此时归一化值为 1，X 取最小值时分子最小为 0，此时归一化值为最小值 0，归一化处理后，数字被压缩在[0,1]之间，逆向化的实际意义为将数字压缩在[0,1]之间，并且调换数字的相对大小意义，比如原始数字越大越差（类似负债），处理后数字变成越大越好。因而在实际研究中，逆向化指标进行逆向化处理后，就会变成正向指标。

$$\text{逆向化:}\frac{x_{\text{Max}}-x}{x_{\text{Max}}-x_{\text{Min}}}$$

适度化时，k 值为一个输入参数值，比如 k 值=1，其意义为数字越接近于 1，适度化后数字越大，适度化处理后数字均小于等于 0，但越接近于 0 说明其离 k 值越近。

$$\text{适度化:}-|x-k|$$

区间化时，a 值和 b 值均为输入参数值，比如 a 值=1 且 b 值=2，其意义将数据压缩在[1,2]之间，区间化是归一化的通用化公式，将数字压缩在设置的范围内，并且保持其数字的相对大小意义不变化。

$$\text{区间化:}a+(b-a)\times\frac{(x-x_{\text{Min}})}{x_{\text{Max}}-x_{\text{Min}}}$$

初值化时，分母x_0为原始数字的第 1 个值，通常其意义为比如 2000 年的 GDP，其余 2000 后的 GDP 数据与 2000 年进行对比，处理后意义为 2000 年 GDP 的倍数，初值化通常仅针对全部均大于 0 的数字。

$$\text{初值化:}\frac{x}{x_0},x_0\text{表示初值}$$

最小值化时，分母x_{Min}为原始数字最小值，其处理后数字的意义为最小值的多少倍，最小值化通常仅针对全部均大于 0 的数字。

$$\text{最小值化:}\frac{x}{x_{Min}}$$

最大值化时，分母x_{Max}为原始数字最大值，其处理后数字的意义为最大值的多少倍，最大值化通常仅针对全部均大于 0 的数字。

$$\text{最大值化:}\frac{x}{x_{Max}}$$

求和归一化时，分母是所有数字求和，其处理后数字的意义是各数字的占比，求和归一化通常仅针对全部均大于 0 的数字。

$$\text{求和归一化:}\frac{x}{\sum_{i=1}^{n}x_i}$$

平方和归一化时，分母是所有数字平方之和然后开根号，平方和归一化的意义为得到数字的相对大小，处理后数字一定小于等于 1，平方和归一化通常仅针对大于 0 的数字。

$$\text{平方和归一化:}\frac{x}{\sqrt{\sum_{i=1}^{n}x_i^2}}$$

固定值化时，FixedValue 是一个输入参数值，比如其为 10，下式中分母为一固定值，其表示所有数字离 10 的最远距离。固定值化的实际意义为离 10 的相对距离（处理后数字越大越接近，数字越小越远离），处理后数字介于[0,1]之间，0 代表远离 10，1 表示刚好为 10。

$$固定值化: x_i = 1 - \frac{|\ x_i - \text{FixedValue}\ |}{max\ |\ x - \text{FixedValue}\ |}$$

偏固定值化时，FixedValue 是一个输入参数值，比如其为 10，上式中分母为一固定值，其表示所有数字离 10 的最远距离。偏固定值化的实际意义为离 10 的相对距离（处理后数字越大越远离，数字越小越接近），处理后数字介于[0,1]之间，0 代表刚好为 10，1 表示远离 10。

$$偏固定值化: x_i = \frac{|\ x_i - \text{FixedValue}\ |}{max\ |\ x - \text{FixedValue}\ |}$$

近区间化时，p 值和 q 值是两个输入参数值，比如 p 值=10，q 值=20，如果数字在[10,20]区间内，那么说明在该区间，处理后为 1。如果不在[10,20]这一区间，处理后数字越大意味着越接近该区间，数字越小意味着越远离该区间，且处理后数字介于[0,1]之间。

$$近区间化: x_i = \begin{cases} 1 - \dfrac{\max(p - x_i, x_i - q)}{\max(p - \min(x), \max(x) - q)} & , x_i \notin [p, q] \\ 1 & , x_i \in [p, q] \end{cases}$$

偏区间化时，p 值和 q 值是两个输入参数值，比如 p 值=10，q 值=20，如果数字在[10,20]区间内，那么说明在该区间，处理后为 0。如果不在[10,20]这一区间，处理后数字越大意味着越远离该区间，数字越小意味着越接近该区间，且处理后数字介于[0,1]之间。

$$偏区间化: x_i = \begin{cases} \dfrac{\max(p - x_i, x_i - q)}{\max(p - \min(x), \max(x) - q)} & , x_i \notin [p, q] \\ 0 & , x_i \in [p, q] \end{cases}$$

科学计算

SPSSAU 还提供常用的科学计算，包括取数据的平方、根号、绝对值、倒数、相反数、三次方或取整等。

汇总处理

比如有 X1,X2 和 X3 共三项时，其编号 1 时对应最大值为 5，最小值为 1，中位数为 2。其实际意义为比如取 3 门课程的最高分，最低分，或者中间分数。

编号	X1	X2	X3	最大值	最小值	中位数	计数
1	1	2	5	5	1	2	1
2	2	3	3	2.5	5	6	0
3	3	4	2	3.5	7	12	0
4	3	5	3	4	8	15	0

如果是‘计数’，比如 X1,X2,X3 共三项，计算这三项中出现某个值（该值为输入参数值）的出现次数，比如计算 X1,X2,X3 共三项中出现 1 的次数，那么编号 1 时 3 个数字分别是 1、2 和 5，出现 1 的次数为 1。其实际意义为比如对错题，分别使用数字 1 表示正确数字 0 表示错误，那么统计 1 的次数即为统计选对的次数为多少。

日期相关处理

日期处理包括取出日期数据的基本信息，包括年、月、日等，还涉及到日期相减、日期数据滞后处理、差分处理和季节差分等，如表所示。

项	说明
日期处理	取出日期数据信息，包括年、月、日、周或者季度
日期相减	两个日期数据相减
滞后处理	将时间序列进行滞后处理
差分处理	将时间序列进行滞后处理
季节差分	将时间序列进行季节性差分处理

编号	日期 1	日期 2	日期 1-年	日期 1-月	日期 1-日期 2	时间序列数据	滞后 1 阶	差分 1 阶
1	2023-3-1	2022-11-22	2023	3	99	100	null	null
2	2023-2-28	2022-11-23	2023	2	97	98	100	-2
3	2023-2-27	2022-11-24	2023	2	95	95	98	-3
4	2023-2-26	2022-11-25	2023	2	93	97	95	2
5	2023-2-25	2022-11-26	2023	2	91	78	97	-19
6	2023-2-24	2022-11-27	2023	2	89	76	78	-2

针对日期处理，针对日期 1 取其年份和月份数据，见第 4 列和第 5 列。并且日期 1 与日期 2 相减后得到两个日期的差值天数，见第 6 列。

针对时间序列数据（第 7 列），其滞后 1 阶为第 8 列，即当前日期时数据是上 1 个日期的数据，滞后 2 阶指当前日期时数据是上上个日期的数据。差分 1 阶（第 9 列），其指当前日期数据减去上 1 个日期的数据，比如表格中编号为 2 时对应 98 - 100= -2。

如果是季节周期性数据，可先设置季节周期值后进行差分处理，其原理与普通差分处理类似，比如季节差分 1 阶，季节周期值为 4（比如一年 4 个季度则季节周期值为 4），那么差分是指当前日期数据 – 前 1 个季节周期值对应的数据。

其它

SPSSAU 生成变量中还提供包括样本编号、Box-Cox 变换，秩、缩尾处理、截尾处理等共计 10 项功能，如表所示。

项	说明
样本编号	比如 100 个样本，编号为 1 到 100，共提供顺序编号和随机编号两种方式
Box-Cox 变换	非正态数据转换处理的一种方式
秩	数据的秩
缩尾处理	一种异常值的处理方式
截尾处理	一种异常值的处理方式
Johnson 转换	非正态数据转换处理的一种方式
排名	数据的排名情况
相除	两个数据相除
相减	两个数据相减
非负平移	出于非负数，将数据平移一个单位，使其全部大于 0

比如共有 100 个样本，针对该样本设置一个编号，可以从 1 到 100 顺序递增（顺序编号），也可以随机编号 1 到 100，比如实际研究中仅希望分析前 50 个样本，此时可先设置样本编号，然后复选出编号小于 50 的样本进行分析。

秩是指数据的排名，其与‘排名’功能非常类似，但区别在于如果有几个相同的排名时，秩会取排名平均值，比如前 3 个的数字均相同，那么排名上均为 1（升序时），而秩会取排名的平均值即 0.3333333。

通常在计量研究时会对异常数据进行处理，缩尾处理包括双向缩尾、上侧单向缩尾和下侧单向缩尾，双向缩尾指比如将是小于 2.5%分位数值设置为 2.5%分位数时值，大于 97.5%的值设置为 97.5%分位数的值。如果是上侧单向缩尾，此只会将大于 97.5%的值设置为 97.5%分位数的值，如果是下侧单向缩尾，此只会将小于 2.5%分位数值设置为 2.5%分位数时值（注：此处参数默认值为 0.05 即 5%，研究人员可对其进行设置，系统处理时以小于该参数值/2，或大于 1-该参数值/2 分别作为上侧或下侧的临界值标准）。

相除是指将两项数据进行相除，选择被除数和除数项后即可，类似还有相减功能。非负平移指数据出现小于等于 0 时，则全部加上一个‘平移值’，该‘平移值’=数据最小值的绝对值+参数值（注：参数值默认为 0.01），其意义为让数据全部均大

于 0（并且大于等于参数值）。如果数据全部都大于 0 则系统不会进行非负平移。如下表所示：

编号	X1	X2	X1 非负平移 0.01	X2 非负平移 0.01
1	-1	2	0.01	2
2	2	3	3.01	3
3	3	4	4.01	4
4	3	5	4.01	5

表格中 X1 中出现-1 即小于等于 0 的数字，且 X1 时最小值为-1，其绝对值为 1，那么平移值=1+0.01（0.01 为参数值），此时非负平移后数据见第 4 列。X2 全部均大于 0，因此非负平移并不生效，其处理后数据保持一致（见第 5 列）。