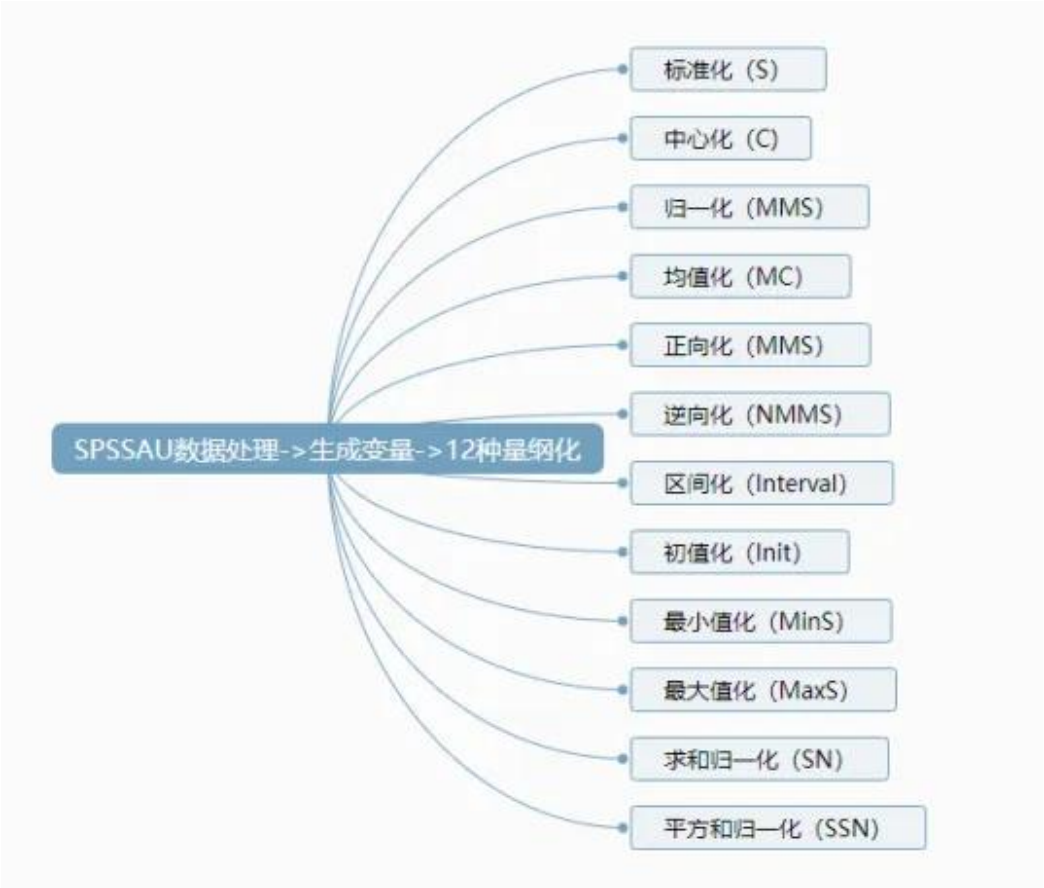
12 种数据量纲化处理方式

在进行数据分析时,数据具有单位是非常常见的,比如说 GDP 可以以亿作为单元,也可以以百万作为单位，那么此时就会出现由于单位问题导致的数字大小问题;这种情况对于分析可能产生影响，因此需要对其进行处理，但是处理的前提是不能失去数字的相对意义，即之前数字越大代表 GDP 越高，处理后的数据也不能失去这个特性，类似这样的处理我们统称为**量纲化**。

也或者计算距离，数字 1 和 2 的距离可以直接相减得到距离值为 1； 另外一组数据为 10000 和 20000，两个数字直接相减得到距离值为 10000。如果说距离数字越大代表距离越远，那么明显的 10000 大于 1，但这种情况仅仅是由于数据单位导致的，而并非实际希望如何，因此就需要进行量纲化处理。

量纲化有很多种方式，但具体应该使用那一种方式，并没有固定的标准，而应该结合数据情况或者研究算法，选择最适合的量纲化处理方式，SPSSAU 共提供 12 种量纲化处理方法，如下图。



1 量纲化基本说明

关于量纲化，其具体的公式计算如下，接下来会逐一说明。

类型	意义	公式
标准化（S）	让数据变成平均值为0，标准差为1	（X - Mean）/ Std
中心化（C）	让数据变成平均值为0	X - Mean
归一化（MMS）	让数据压缩在【0，1】范围内	（X - Min）/ (Max - Min)
均值化（MC）	以平均值作为标准进行对比	X / Mean
正向化（MMS）	让数据压缩在【0，1】范围内	（X - Min）/ (Max - Min)
逆向化（NMMS）	让数据压缩在【0，1】范围内，且数据方向颠倒	（Max - X）/ (Max - Min)
区间化（Interval）	让数据压缩在自己希望的范围内	将数据压缩在a和b之间，默认分别是1和2。a + (b - a) * (X - Min)/(Max - Min)
初值化（Init）	数据除以第1个数字	X / 该列第1个不为空的数据
最小值化（MinS）	以最小值作为标准进行对比	X / Min
最大值化（MaxS）	以最大值作为标准进行对比	X / Max
求和归一化（SN）	数据表达总和的比例	X / Sum（X）
平方和归一化（SSN）	数据表达平方和的比例	X / Sqrt（Sum（X^2））

12 种量纲化类型

备注：表格中，X 表示某数据，Mean 表示平均值，Std 表示标准差；Min 表示最小值，Max 表示最大值，Sum 表示求和，Sqrt 表示开根号。

1）标准化(S)

标准化是一种最为常见的量纲化处理方式。其计算公式为：（X-Mean）/ Std。

此种处理方式会让数据呈现出一种特征，即数据的平均值一定为 0，标准差一定是 1。针对数据进行了压缩大小处理，同时还让数据具有特殊特征（平均值为 0 标准差为 1）。

在很多研究算法中均有使用此种处理，比如聚类分析前一般需要进行标准化处理，也或者因子分析时默认会对数据标准化处理。

比如聚类分析时，其内部算法原理在于距离大小来衡量数据间的聚集关系，因此默认 SPSSAU 会选中进行标准化处理。

除此之外，还有一些特殊的研究方法，比如社会学类进行中介作用，或者调节作用研究时，也可能会对数据进行标准化处理。

2) 中心化(C)

中心化这种量纲处理方式可能在社会科学类研究中使用较多，比如进行中介作用，或者调节作用研究。其计算公式为：X - Mean。

此种处理方式会让数据呈现出一种特征，即数据的平均值一定为0。针对数据进行了压缩大小处理，同时还让数据具有特殊特征（平均值为0）。

平均值为0是一种特殊情况，比如在社会学研究中就偏好此种量纲处理方式，调节作用研究时可能会进行简单斜率分析，那么平均值为0表示中间状态，平均值加上一个标准差表示高水平状态；也或者平均值减一个标准差表示低水平状态。

3) 归一化(MMS)

归一化的目的是让数据压缩在【0，1】范围内，包括两个边界数字0和数字1；其计算公式为（X - Min）/（Max - Min）。

当某数据刚好为最小值时，则归一化后为0；如果数据刚好为最大值时，则归一化后为1。

归一化也是一种常见的量纲处理方式，可以让所有的数据均压缩在【0，1】范围内，让数据之间的数理单位保持一致。

4) 均值化(MC)

均值化在综合评价时有可能使用，比如进行灰色关联法研究时就常用此种处理方式；其计算公式为X / Mean，即以平均值作为单位，全部数据均去除以平均值。

需要特别说明一点是，此种处理方式有个前提，即所有的数据均应该大于0，否则可能就不适合用此种量纲方式。

5) 正向化(MMS)

正向化的目的是对正向指标保持正向且量纲化，什么意思呢。比如这样一些指标GDP增长率、科研产出数量、失业率共3个指标；明显的，GDP增长率、科研产出数量是数字越大越好，而失业率是数字越小越好。

正向化的目的就是让数字越大越好的意思，而且同时其还让数据压缩在【0，1】范围内即进行了量纲处理。其计算公式为（X - Min）/（Max - Min）。

当某数据刚好为最小值时，则归一化后为0；如果数据刚好为最大值时，则归一化后为1。

正向化和归一化的公式刚好完全相等,但正向化强调让数字保持越大越好的特性且对数据单位压缩，而归一化仅强调数字压缩在【0，1】之间。

正向化的使用情况为：当指标中有正向指标，又有负向指标时；此时使用正向化让正向指标全部量纲化；也或者指标全部都是正向指标，让所有正向指标都量纲化处理。

6) 逆向化(NMMS)

逆向化的目的是对逆向指标正向且量纲化，什么意思呢。比如这样一些指标 GDP 增长率、科研产出数量、失业率共 3 个指标；明显的，GDP 增长率、科研产出数量是数字越大越好，而失业率是数字越小越好。

逆向化的目的就是让数字越小越好的意思，而且同时其还让数据压缩在【0，1】范围内即进行了量纲处理。其计算公式为（Max - X）/（Max - Min)。

从公式就可以看出，分母永远是大于 0，随着 X 的增大，分子会越来越小，那么就对逆向指标逆向化处理之后就会得到一个这样的特征，即数字越大越好（数字越大时，其实 X 是越小）。

相当于将逆向指标逆向化后，新的数据为数字越大越好，这样便于进行方向的统一,尤其是在指标同时出现正向指标和逆向指标时,针对逆向指标进行逆向处理,是非常常见的处理方式。

7) 区间化(Interval)

区间化的目的是让数据压缩在【a，b】范围内，a 和 b 是自己希望的区间值，如果 a=0,b=1，那么其实就是一种特殊情况即归一化；其计算公式为 a + (b - a) * (X - Min)/(Max - Min)。

此公式会让数据永远的保持在【a,b】之间，SPSSAU 默认 a 为 1，b 为 2，即将数据压缩在【1，2】之间，当然研究者根据需要进行设置即可。它的目的仅仅是对数据进行压缩在固定的区间，保持数据数理单位的一致性。

8) 初值化(Init)

初值化在综合评价时有可能使用,比如进行灰色关联法研究时就常用此种处理方式；其计算公式为 X / 该列第 1 个不为空的数据，即以数据中第 1 个不为空的数据作为参照标准，其余的数据全部去除以该值。

比如说 2000，2001，2002，2003，一直到 2020 共计 21 年的 GDP 数据，第 1 个数据就是 2000 年的 GDP，所有的数据都去除以 2000 年的 GDP，相当于以 2000 年 GDP 作为参照标准，所有数据全部除以 2000 年的 GDP（包括 2000 年 GDP 除以自己得到数字 1）。

一般来说，初值化这种处理方式适用于有着一种趋势或规律性的数据，比如上述2000~2020年的GDP等，而且数据正常情况下都是全部大于0，因为出现负数，通常会失去其特定意义。

9）最小值化(MinS)

最小值化，其目的是让最小值作为参照标准，所有的数据全部除以最小值；其计算公式为X／Min，即以最小值作为单位，全部数据全部去除以最小值。需要特别说明一点是，此种处理方式时一般都是要求数据全部大于0，否则可能就不适合用此种量纲方式。

10）最大值化(MaxS)

最大值化，其目的是让最大值作为参照标准，所有的数据全部除以最大值；其计算公式为X／Max，即以最大值作为单位，全部数据全部去除以最大值。需要特别说明一点是，此种处理方式时一般都是要求数据全部大于0，否则可能就不适合用此种量纲方式。

11）求和归一化(SN)

求和归一化，其目的是让‘求和值’作为参照标准，所有的数据全部除以求和值，得到的数据相当于为求和的占比；其计算公式为X／Sum（X），即以所有数据的‘求和值’作为单位，全部数据全部去除以‘求和值’。

需要特别说明一点是，此种处理方式时一般都是要求数据全部大于0，否则可能就不适合用此种量纲方式。TOPSIS法的时候使用此种处理方式较多。

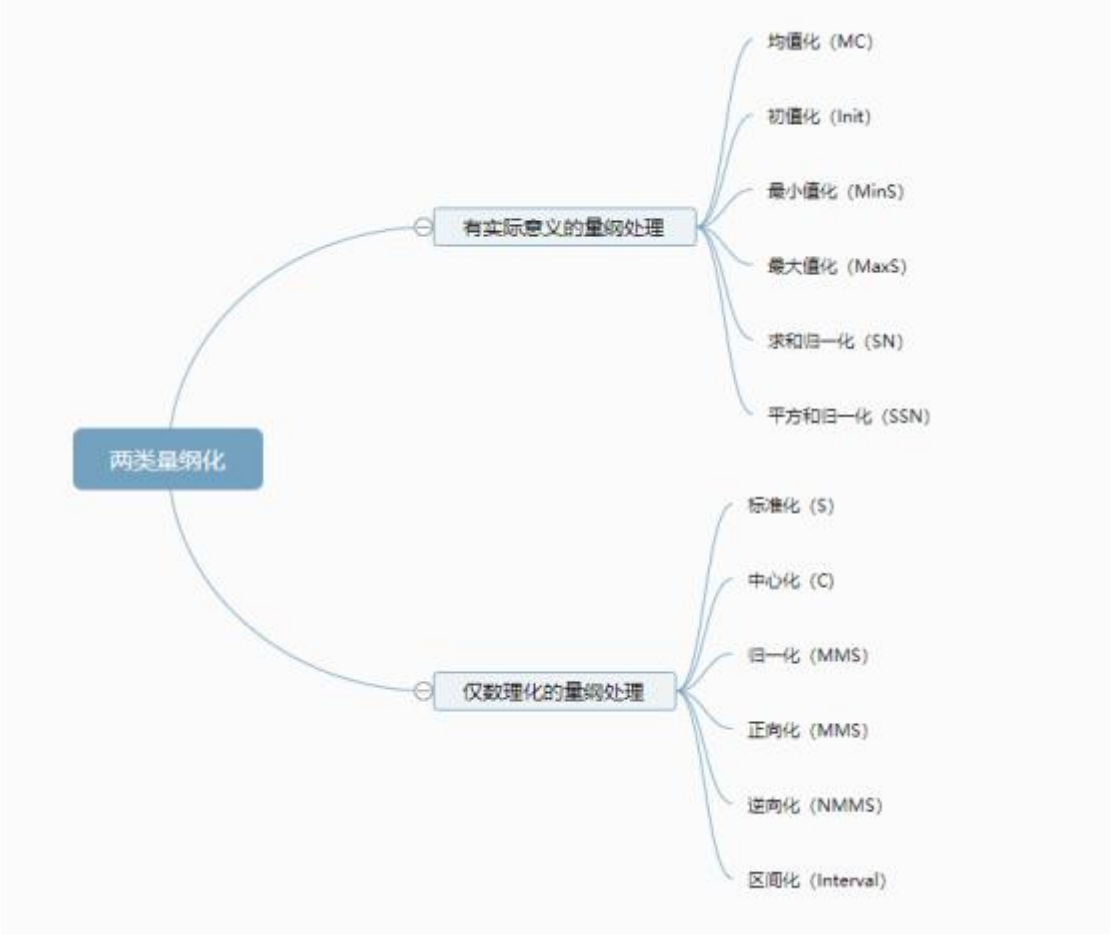
12）平方和归一化(SSN)

平方和归一化，其目的是让‘平方和值’作为参照标准，所有的数据全部除以平方和值，得到的数据相当于为平方和的占比；其计算公式为X／Sqrt(Sum(X^2))，即以所有数据的‘平方和值’作为单位，全部数据全部去除以‘平方和值’。

需要特别说明一点是，此种处理方式时一般都是要求数据全部大于0，否则可能就不适合用此种量纲方式。TOPSIS法的时候使用此种处理方式较多。

3 量纲化如何使用

量纲化按是否具有实际意义可分为两类，一类是量纲处理方式有着一定的实际意义，另一类是仅数理角度的量纲处理方式；如下图：



量纲化的目的是将数据进行量纲单位统一化，有的量纲化具有实际意义，比如均值化，初值化，最小值化，最大值化，和求和归一化，平方和归一化共 6 种。分别代表数据除以平均值，数据除以第 1 个数，数据除以最小值，数据除以最大值，数据除以求和值，数据除以平方和值。相当于说，它们都找到一个参照标准项，然后所有数据去除以参照标准项。此 6 种方式的特点在于，一般要求数据全部都大于 0，如果出现小于 0 或者等于 0 就有可能出问题，比如刚好分母为 0，那么就出现无法相除。

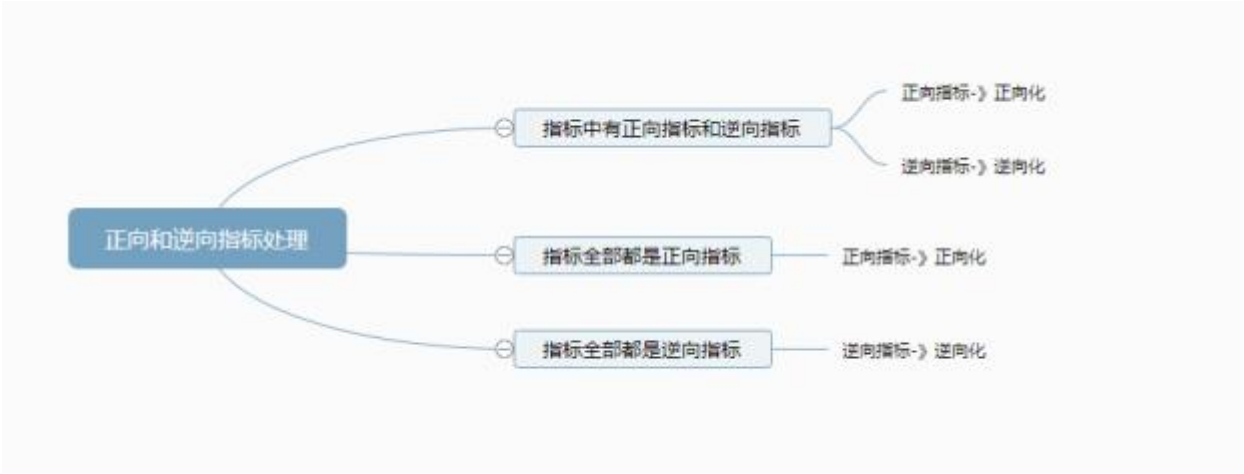
除此之外，仅数理化的量纲处理，包括标准化，中心化，归一化，正向化，逆向化，区间化，均在于让数据保持在一定的区间范围内，而且处理后带有一定的数理特征，比如标准化后数据的平均值为 0 标准差为 1；中心化后数据平均值为 0；归一化后数据最小为 0 最大为 1；正向化后数据最小为 0 最大为 1；逆向化后数据最小为 0 最大为 1；区间化是研究者自行设定处理后数据压缩在对应的范围内。

在研究时具体应该使用那一种处理方式呢，其实并没有固定的要求，而是结合实际情况或者实际研究进行。比如社会学类的中介作用和调节作用偏好于使用中心化或标准化这种处理方式；聚类分析或者因子分析等使用默认会使用标准化；综合评价时比如灰色关联法偏好于使用均值化或初值化；TOPSIS 法时偏好于使用求和归一化或者平方和归一化。如果想对数据的指标方向进行统一，那么就会使用正向化或者逆向化。

如果单独想对数据量纲进行处理（且没有分析方法上的常用习惯），那么**通常默认是使用标准化或者归一化最多**，标准化直接把数据压缩且数据有一种特质即平均值为 0 标准差为 1 的特质；归一化把数据压缩在【0，1】之间。也或者使用中心化让数据有一种特质即平均值为 0。

如果数据中有负数，正常情况下不能使用‘有实际意义的量纲处理’即均值化，初值化，最小值化，最大值化，求和归一化，平方和归一化。

特别说明，正向化和逆向化这两种处理方式，其目的有 2 个，一是对数据进行量纲单位处理，最终让数据压缩在【0，1】之间。除此之外，其还可以对正向或负向指标进行方向上的统一；如果数据包括正向和逆向指标，那么正向指标进行正向化处理，负向指标进行负向化处理，最终让所有的指标都压缩在【0，1】之间，而且都让指标有一个物质即数字越大越好。如果说指标全部都是正向指标那么全部正向化即可，正向化后数字还是越大越好；如果说指标全部都是逆向指标那么全部逆向化即可，逆向化后数字就代表越大越好。


正向和逆向指标处理
指标中有正向指标和逆向指标
正向指标-》正向化
逆向指标-》逆向化
指标全部都是正向指标
正向指标-》正向化
指标全部都是逆向指标
逆向指标-》逆向化