异常值

目录

如果在探索分析时发现数据中有着异常值，可经【数据处理】→【异常值】进行异常值设置和处理。异常值设置包括‘判断标准’和‘异常值处理’共两项，如下表所示。

项	说明
判断标准	设定异常值的标准，比如身高大于 2.2 米
异常值处理	针对符合异常值，进行对应的处理

判断标准

关于异常值判断标准，比如在探索分析时发现身高大于 2.2 米，那么可在判断标准时设置‘数字>2.2’，判断标准的设置上还包括‘缺失数字’，其指原始数据中某项为 null 值。与此同时，判断标准还包括等于某个数字、小于某个数字或者数字偏离标准差幅度（通常是 3 倍或者 2 倍标准差，可见 2.2.3 节内容）。判断标准共有 5 个条件，5 个条件是或者关系，选择任意一个或多个判断标准时，只要选中的任意一个判断标准满足即可。如下图：

异常值处理

设置好异常值判断标准之后，接着需要设置如何进行异常值处理，如下表所示，异常值处理共分为三种方式，分别是‘设置为 Null’，‘填补’和‘插值’，三者为互斥关系，只能选择其中任意一种。

处理方式	说明
设置为 Null	将满足条件的异常值设置为 Null 值

填补	包括平均值、中位数、众数、随机数、数字 0 和自定义共 6 种方式
插值	包括线性插值和‘该点线性趋势插值’两种方式

‘设置为 Null’这种处理方式使用较多，直接将满足条件的异常值设置为 Null 值，但此种方式处理后，分析的有效样本量可能会减少，因而其适用于样本量相对较大时使用。

如果是填补法，其包括 6 种方式，其中平均值、中位数或众数填补，其指将除异常值外的数据进行取平均值（中位数或众数），然后将异常值全部替换成该平均值（中位数或众数）。如果是随机数填补，系统会将满足条件的异常值替换成一个随机数字。‘数字 0’填补指系统将满足条件的异常值替换成数字 0，自定义填补指系统将满足条件的异常值替换成‘主动设置的一个具体数字’。

插值法时，共包括两种，分别是线性插值和‘该点线性趋势插值’两种，具体下表所示。

编号	原始数据	线性插值	该点线性趋势
1	1	1	1
2	3	3	3
3	4	4	4
4	异常值	5	5.46511
5	6	6	6
6	9	9	9

线性插值指：将挨着异常值最近的前面 1 个点和后面 1 个点，比如表格中编号为 4 这个异常值，其前面一个点编号为 3 数据为 4 即（3，4），其后面一个点编号为 5 数据为 6 即（5，6），将该两个点连成一条线时会得到坐标公式，然后代入异常值的编号最终计算得到线性插值。比如表格中（3，4)和（5，6）两个点连成直线时，该直线的坐标公式为：y = x + 1，代入异常值对应的编号 4，得到线性插值的具体值=4+1=5。另外，编号指原始数据的编号，比如有 100 个样本，编号则从 1、2、3 一直递增到 100。

如果是‘该点线性趋势’，其指将除异常值外的其它正常数据进行线性回归拟合（其中 Y 为原始数据，X 是编号），在得到线性回归方程公式后（关于线性回归，可参考第 5 章内容），然后将异常值对应的编号作为 X 代入，计算得到的 Y 值即为‘该点线性趋势’。除异常值外，共有 5 行数据，该编号作为 X，原始数据作为 Y，进行线性回归拟合得到线性回归方程公式即：Y = -0.30233 + 1.44186 × X，代入异常值时编号为 4，即‘该点线性趋势’= -0.30233 + 1.44186 × 4 = 5.46511。