抽样节点使用说明

作者：张青松

目录

[1. 抽样 2](#_Toc430940383)

[2. 抽样方法 2](#_Toc430940384)

[2.1. 简单抽样 2](#_Toc430940385)

[2.2. Bootstrap抽样 3](#_Toc430940386)

[2.3. 等容量抽样 3](#_Toc430940387)

[3. 抽样节点设置和使用 3](#_Toc430940388)

# 抽样

抽样又称取样。是指使用统计学方法，从待研究的样本总体中抽取一部分样本单位，抽取的样本需要对总体具有充分的代表性。

抽样的目的是从对被抽取样本的分析、研究结果，来估计和推断总体的特征，是科学实验，质量检验、社会调查普遍采用的一种经济有效的研究方法。

抽样方法从是否放回样本可以分为：可放回抽样和不放回抽样。

# 抽样方法

在Datastudio中的抽样节点中，将抽样分为三个大类，分别是：简单抽样、Bootstrap、等容量抽样。其中简单抽样和等容量抽样是不放回抽样，Bootstrao是可放回抽样。

## 简单抽样

在简单抽样设置面板最上边，设置抽样的数量，可以选择“数量”，表示从样本总体中抽取固定数目个样本；也可以选择“百分比”，表示根据样本总体的数量的百分比，抽取一定数目的样本。

简单抽样的抽样方法包括：

1. 从第一个抽样：根据要抽取样本数，从头按顺序抽出样本。
2. 线性抽样：按固定间隔从样本中抽取给定数目的样本。
3. 随机抽样：随机从总体中抽取给定数目的样本数。
4. 分层抽样：先根据指定的列将样本总体分成若干类，在根据每一类在样本中占的比例，从该类中抽出一定数目的样本。最后将所有类抽出来的样本合并在一起，作为抽样结果。

对于抽样过程中的随机性，可以使用随机种子来控制，当使用了相同的随机种子，抽样结果相同。Bootstrap和等容量抽样中的随机种子也相同。

## Bootstrap抽样

Bootstrap抽样，是可放回抽样，通过设置面板设置是否添加数据的原始行号，和该行被抽中的次数。

## 等容量抽样

等容量抽样是指根据指定的列，将数据分成若干类，并找到所有分类中数据行数最少的数，然后从每个类中都抽出最少分类的个数，组成样本。

等容量节点中，有两种抽样：

1. 精确抽样：严格根据最小行数在每个分类中抽取样本。
2. 近似抽样：在每个分类中，根据最小行数的在附近随机上下波动，计算得到抽样个数。根据大树定理，当分类数较大时，总样本数趋近与精确抽样的样本数。

# 抽样节点设置和使用

抽样节点是一个数据处理的节点，输出结果是输入数据的一个子集。在工作流中添加抽样节点，并连接输入数据。



图1 抽样节点工作流图

打开设置界面，根据抽样需求设置抽样参数。

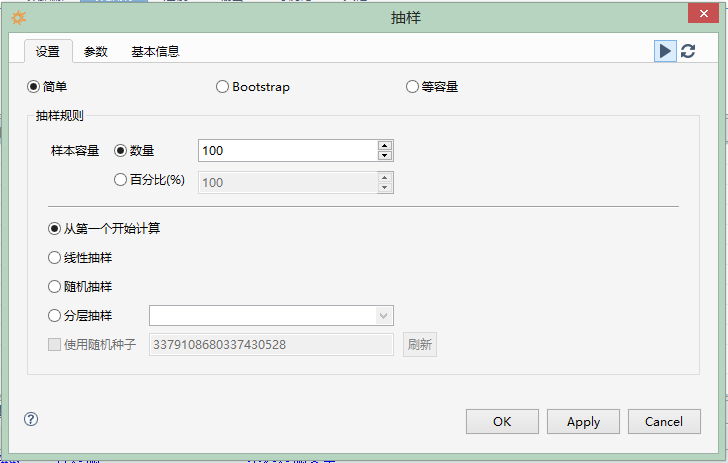


图2 抽样节点设置面板

设置完成后，执行节点，右键查看抽样结果。

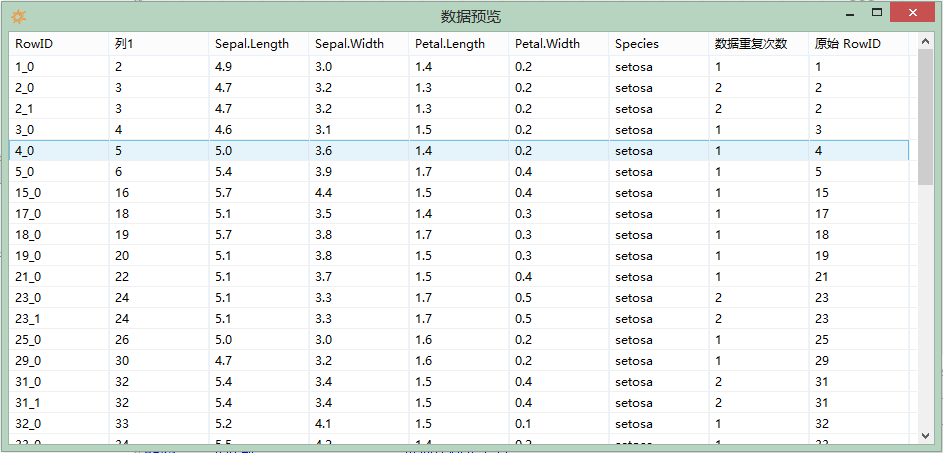


图3 Bootstrap抽样结果

针对Bootstrap的抽样结果，说明。第一列RowID下划线之前的部分（下划线可是通过参数中的RowID分隔符设置），是该数据行原始的RowID，下划线后面的部分是在可放回抽样中被抽取出来的次序，从0开始。倒数第二列的“重复次数”是指，该行对应的原始数据被抽出的次数。最后一列“原始RowID”和第一列RowID下划线之前的部分相同。