数据准备相关概念总结

1. 什么是样本个数、变量个数和数据取值?

样本个数: 就是数据表的行数，由于每一行数据也叫做一条记录，所以样本的个数也可以说是数据表的记录数。

变量个数：调查对象的特征或属性称为变量，由于第一列为调查对象，所以除了第一列外的剩余列数称为变量个数。很多情况下，我们可以直接用数据表的列数来表示变量的个数。

数据取值：每个字段的取值即为数据取值。

例如:

姓名	性别	年龄	学习成绩	课外书种类
张三	男	19	优	小说
李四	女	20	良	传记
王五	男	20	优	漫画
赵六	女	20	良	小说
孙七	男	21	差	科普

想一想：样本、变量个数是不是越多越好?

样本个数太少会影响结果的解释程度，样本太多则增加统计难度和数据剖析困难。

同样变量个数也并非越多越好，变量个数太少，不能全面的反映问题，但是变量个数如果太多，则会造成数据维度高，使分析时间过长或无法收敛。

2. 数据可以分为哪些类型?

类别型数据：用来描述性质或者特征的，也称为定性数据。

数值型数据：用来描述数量的，涉及到的是数字，也叫做定量数据。

例如：

类别型数据	数值型数据
课外书种类	年龄
小说	19
传记	20
漫画	20
小说	20
科普	21

3. 如何描述数据的质量?

1）集中趋势：均值、中位数、众数

均值：通将所有数据加起来，然后除以数字个数。

中位数：将数据按大小排列后，其中间值被称作中位数。如果数据的数量是偶数，中位数就等于中间两个数值的均值。

众数：数据中出现次数最多的数值称为众数。

2）离散程度：方差、标准差

方差：数值与均值之差的平方数的均值。方差越小，数据的离散程度越小。

标准差：数值与均值距离的均值。为了求出标准差，先计算方差，然后取其平方根。标准差越小，数值离均值距离越近，数据的离散程度越小。