统计基础---描述数据集的参数

对于一个数据集，常使用以下参数从不同角度描述数据集：

均值(mean): 数据集的平均值，易受异常值的影响

众数(mode): 数据集出现次数最多的数据

中位数(median):数据集排序后，中间位置的数据

值域(range):数据集中最大值减最小值。如果存在异常数据，比如特别大或者特别小，将会影响值域。通常采用四分位差(IQR): 数据集排序后，去掉前 25%，去掉后 25%，剩余数据最小值记为 Q1，最大值记为 Q3，IQR=Q3-Q1。判定一个数据是否为异常值：data>Q3+1.5*IQR 或者 data<Q1-1.5*IQR。由计算过程可以看出，IQR 并未考虑所有数据，即使两个完全不同的数据集，也可能会有相同的 IQR。

平方和(sum of squares): $ss = \sum(x_i - \bar{x})^2$

平均平方偏差(均方差，average squared deviation)：variance=$\frac{\sum(x_i - \bar{x})^2}{n}$

标准差(standard deviation):$\sigma = \sqrt[2]{\frac{\sum(x_i - \bar{x})^2}{n}}$

贝塞尔校正(Bessel's Correction)：上面的均方差和标准差计算公式是在计算总体的时候用的，对于抽样样本，$s = \sqrt[2]{\frac{\sum(x_i - \bar{x})^2}{n-1}}$， 用 s 表示更正后的标准差