数据的基本统计描述

中心趋势度量：均值、中位数和众数

1.1 均值（mean）

令 $x_1, x_2, \ldots, x_N$ 为集合 X 的 N 个观测值，则该集合的均值为：

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

均值是描述数据集中心的较好方法，但对极端值很敏感，如一个班的考试平均成绩可能被少数很低的成绩拉低一些。为抵消少数极端值的影响，可采用丢弃高低极值后的均值。

1.2 中位数（median）

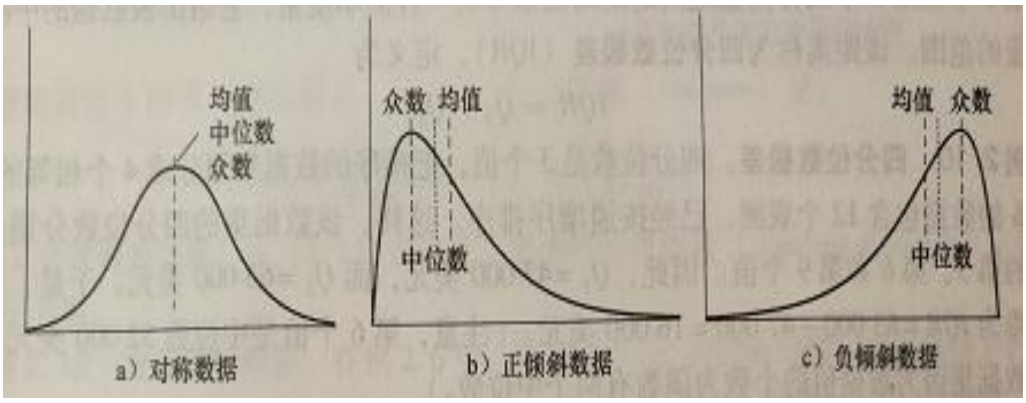
对于非对称数据，数据中心的更好度量是中位数。

对于一个数据集，按递增序排序，若有奇数个观测值，中位数就是中间值，若有偶数个观测值，一般约定，最中间两个值的平均值为中位数

1.3 众数（mode）

数据集中的众数是一组数据中出现最频繁的值，可以对应多个不同的值。

对于只有一个众数的数据集，其观测值分布如下所示：



2. 度量数据散布：四分位数、方差、标准差

2.1 四分位数（quartile）

四分位数是 3 个数据点，它们把按增序排列后的一组数据划分成 4 个相等的部分，使得每部分表示数据分布的四分之一。

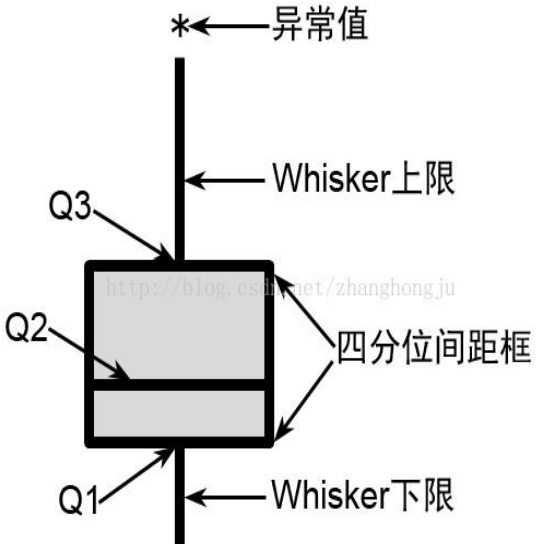
四分位数给出分布的中心、散布和形状的某种指示。第 1 个四分位数（Q1）

是第 25 个百分位数，它砍掉数据的最低的 25%。第 2 个四分位数（Q2）是第 50 个百分位数，作为中位数，它给出数据分布的中心。第 3 个四分位数（Q3）是第 75 个百分位数，它砍掉数据的最低的 75%（或最高的 25%）

四分位数极差：IQR = Q3 － Q1

离群点（或者称异常值），通常是在第 3 个分位数之上或第 1 个四分位数之下至少 1.5*IQR 处的值。

2.2 箱线图（boxplot）或者称箱须图（Box-whisker Plot）是对数据分布的直观表示：



- 箱体的顶部线条是第三个四分位数的位置，即 Q3，表示有 75%的数据小于等于此值。底部线条是第一四分位数的位置，即 Q1，表示有 25%的数据小于此值。整个箱体代表的是数据集中 50%（即 75%-25%）的数据。
- Q2 是数据中位数的位置。
- 箱外两条线称为胡须（Whisker）延伸到最小和最大观测值。
- 对于异常值，在箱线图中：仅当最高和最低观测值超过四分位数不到 1.5*IQR 时，胡须扩展到它们。否则，胡须在四分位数的 1.5*IQR 处终止，剩下的异常值使用星号“*”表示。

2.3 方差和标准差

数据集 X 的 N 个观测值 $x_1, x_2, \ldots, x_N$ 的方差是：

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

其中，$\bar{x}$ 是观测的均值，σ 是标准差，是方差的平方根。

低标准差意味着数据观测趋向于非常靠近均值，高标准差表示数据散布在一个大的值域中。

参考：《数据挖掘：概念与技术》