# 方差和数据的分散性

## 前言

数据间的各种距离，可达到计算数据的分散性和变异性。变异性提供了对一个分布中的数据分散开或聚集在一起的程度的数量测量。

对于均值，中位数，众数单纯的指出数据的中心，但是这样还是不够，并没有的整体的指出的数据频率的情况及稳定性。

## 数据分散性

### 全距

通过计算全局，可以获取数据的分散情况。全局值数据的分散范围，是数据的跨度。计算的方法为：数据集中的最大值减去数据集中的最小值。

* **上界**

最大值称为上界

* **下界**

最小值称为下界

* **异常值对全距的影响**

对于将数据进行排序，在数据的两端的值中可能包含异常时，于是全距的值将会受到影响。由于全距的计算中及其的简单，所以无法很好的进行对数据的分散性进行分析（主要受异常值的影响）。

### 四分位距

为了消除异常值中对全距的影响，可以使用四分位距剔除异常值。

* **计算的方法**
  + 对整份数据进行排序，从小到大进行排序
  + 将数据进行4等分，即每一份的数据为包含4分之一的数据。
  + 求取上下四分位距的值和四分位距，求取中位数
* **第一四分位数 (Q1)**

又称“较小四分位数”，等于该样本中所有数值由小到大排列后第25%的数字。

* **第二四分位数 (Q2)，又称“**[**中位数**](https://baike.baidu.com/item/%E4%B8%AD%E4%BD%8D%E6%95%B0)**”**

等于该样本中所有数值由小到大排列后第50%的数字。

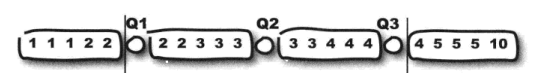
* **第三四分位数 (Q3)**

又称“较大四分位数”，等于该样本中所有数值由小到大排列后第75%的数字。

* [**四分位距**](https://baike.baidu.com/item/%E5%9B%9B%E5%88%86%E4%BD%8D%E8%B7%9D)

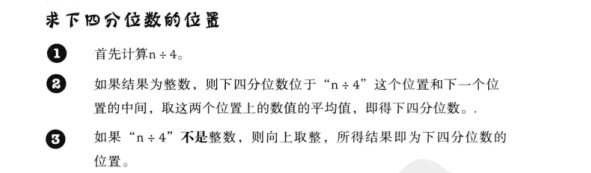
第三四分位数与第一四分位数的差距又称[四分位距](https://baike.baidu.com/item/%E5%9B%9B%E5%88%86%E4%BD%8D%E8%B7%9D" \t "_blank)（InterQuartile Range,IQR）。

**示例图**

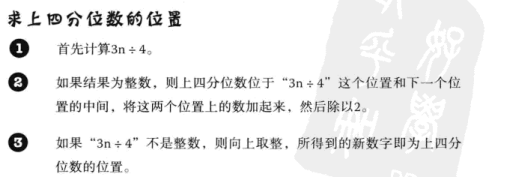


### 算法

* **求解下四分位距**



* **求解上四分位距**

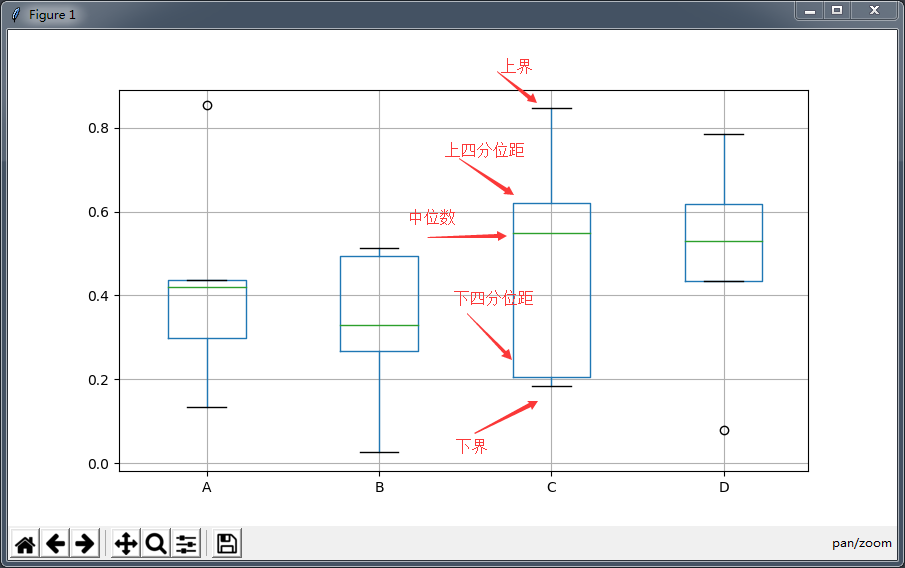


* **求解四分位距**

**对应上图中 使用Q3计算得到的数减去 Q1处计算得到的数， 即Q3 – Q1。**

### 使用箱线图表示距

在python中使用matplotlib得到的箱线图（图中的圆点表示的是异常值）



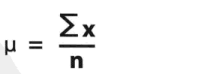
## 数据的变异性

通过数据的全距和四分位距得到数据的分散性，这样也就是可以反应出一些稳定性。但是还是不够，虽然是已经整体上对数据进行了分析，但是无法体现出数据的频率，以及对数据变异程度的分析，数据距离数据中心的情况。

### 使用方差

方差衡量[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)或一组数据时离散程度的度量，概率论中方差用来度量[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)和其[数学期望](https://baike.baidu.com/item/%E6%95%B0%E5%AD%A6%E6%9C%9F%E6%9C%9B)（即[均值](https://baike.baidu.com/item/%E5%9D%87%E5%80%BC)）之间的偏离程度。在计算方差中，一个较为重要的计算量是均值（μ）。

* **对于计算中的整体的思路**
  + **求解均值（**μ**）**

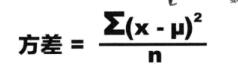


* + **求平均距**

**即各个值到均值的距离的和的平均值，每个值和均值的差计算得到的结果是可正可负的，这样会出现抵消作用，导致平均值的计算不准确。**

* + **求平均距的平方和 – 方差σ**

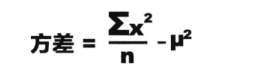
**针对在上述中的问题，为了消除抵消的效果，将每个值和均值的差值进行平方和的平均值，于是这样就得到了 方差 使用符号σ2**



* + **求标准差**

**方差开方得到标准差。**

### 方差速记法



### 标准分

对于不同的数据集间的比较，如何比较它们将的稳定性，对于在这样的情况下进行比较数据是，由于数据集间的均值和标准差都不相同，需要进行一次的装换将它们转化为在相同的情况下。

对于标准分的计算的方法为:

