数据科学基础

Foundations of Data Science

1.3 数据汇总

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

数据汇总

- 集中趋势度量（Central Tendency）
- 离散趋势度量(Variation Tendency)

集中趋势度量

- 集中趋势度量反映的是数据（样本或者总体）的平均水平或数据的中心值。
- 然而“平均”这个词经常在不同场合被混淆。对于“平均”的不同理解，往往导致不同的计算结果。

例如，假如A班15位 毕业生的就业年薪从低到高排列如表

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
年薪	4.5	4.8	5	5	5.2	5.5	5.8	5.8	6	6	6	6	6.5	8.4	14

众数

- 众数是一批数据中出现次数最多的那个数值，通常记为 M 。
- 众数不受极值的影响。众数通常用来描述离散型变量，尤其是分类型变量。

年薪	4.5	4.8	5	5.2	5.5	5.8	6	6.5	8.4	14
累计人数	1	1	2	1	1	2	4	1	1	1

众数通常用于定类变量的统计中,对于定序变量、定距变量和定比变量,通常使用中位数和算术平均数表示集中趋势。对于后三种变量类型,另外一种做法就是先分组再求众数。

中位数

- 中位数是将数据从小到大排序后，处在数据序列中间位置的数值。
- 设一批数据经排序后为X_1，…，X_n，则其中位数M_e为

$$M_e = \frac{n+1}{2} \text{ 对应位置的数据}$$

若n为奇数,则中位数就是中点位置, $\frac{n+1}{2}$,对应的观察值。若n为偶数,则中位数就是中点位置, $\frac{n}{2}$和$\frac{n}{2}+1$,的两个观察值的算术平均数。

序 号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
年 薪	4.8	5	5	5.2	5.5	5.8	5.8	6	6	6	6	6.5	8.4	14

四分位数

- 四分位数是一种常用的集中趋势度量。它通常描述数据序列的四等分的描述性量数。
- 四分位数分为第一四分位数Q_1、第二四分位数Q_2和第三四分位数Q_3 。
- 第四分位数Q_1是这样一个数字:它有25%的数据比它小,75%的数据比它大;第二四分位数Q_2是这样一个数字它有50%的数据比它小,50%的数据比它大;第三四分位数Q_3是这样一个数字:它有75%的数据比它小,25%的数据比它大。
- 容易知道, Q_2等价于我们前面定义的中位数。

四分位数

- 设一批数据经排序后为X_1，…，X_n，则其第i四分位数Q_i为
    - Q_i=$[\frac{i(n+1)}{4}]$ 对应位置的数据 (4)
- 在计算四分位数的时候,要分以下三种情况:
    - 若求得的位置恰好是一个整数,则对应位置的观察值就是相应的四分位数。
    - 若求得的位置不是一个整数,则靠近这个数字的位置上的数据为相应的四分位数。即小数位小于0.5,则取左边位置的数据;小数位大于0.5,则取右边位置的数据。
    - 若求得的位置恰好在两个整数的中间,即小数位为0.5，则计算左右位置数据的算术平均数为相应的四分位数。

四分位数

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
年薪	4.5	4.8	5	5	5.2	5.5	5.8	5.8	6	6	6	6	6.5	8.4	14

A班毕业生年薪的三个四分位数分别为：

$$\frac{(15+1)}{4}=4，Q_1=\text{第4位的数据}=5$$

$$\frac{2(15+1)}{4}=8，Q_2=\text{第8位的数据}=5.8$$

$$\frac{3(15+1)}{4}=12，Q_3=\text{第12位的数据}=6$$

四分位数

若A班的数据去掉第一位同学，即由15位同学变成14位同学，则相应的四分位数为：

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14
年薪	4.8	5	5	5.2	5.5	5.8	5.8	6	6	6	6	6.5	8.4	14

N分位数

N下分位数

N上分位数

（算术）平均数

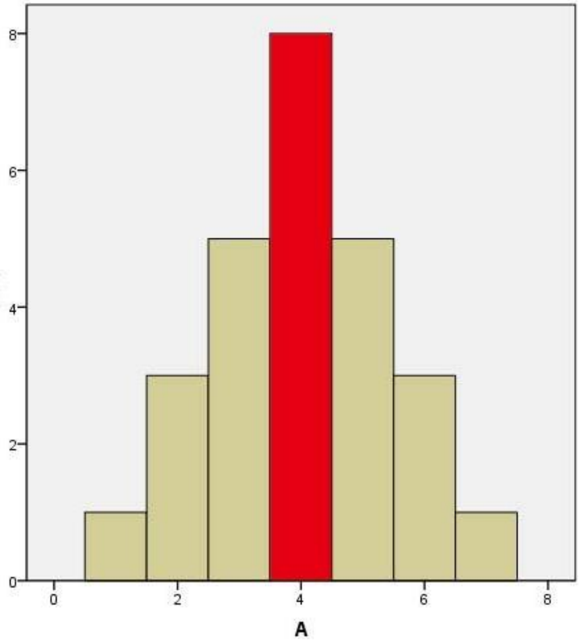
算术平均数，也简称平均数或均值，是最常用的集中趋势度量。它是将所有样本观察值之和除以样本容量。

设一批数据经排序后为X_1，X_2，…，X_n，则其算术平均数 A_n（也记作$\bar{X}$）为

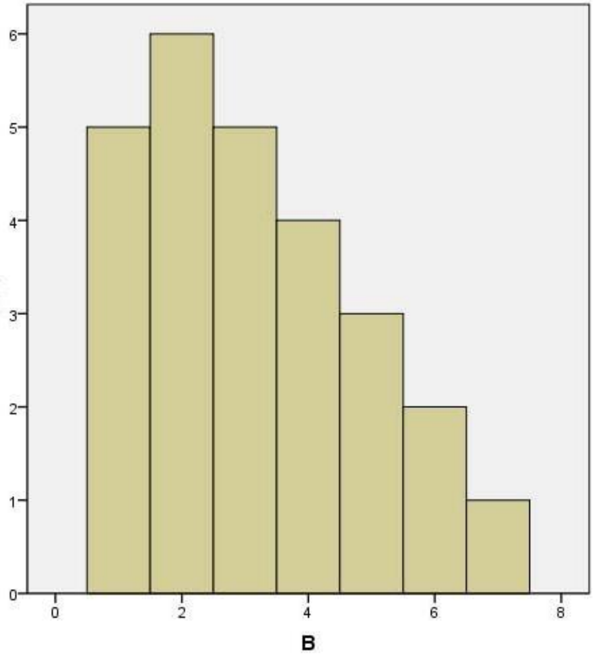
$$A_n = \frac{X_1+X_2+\cdots+X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (5)$$

注：当数据中存在极值时,算术平均数便会歪曲数据反映的信息。此时算术平均数并不是描述这类数据集中趋势的最佳度量。

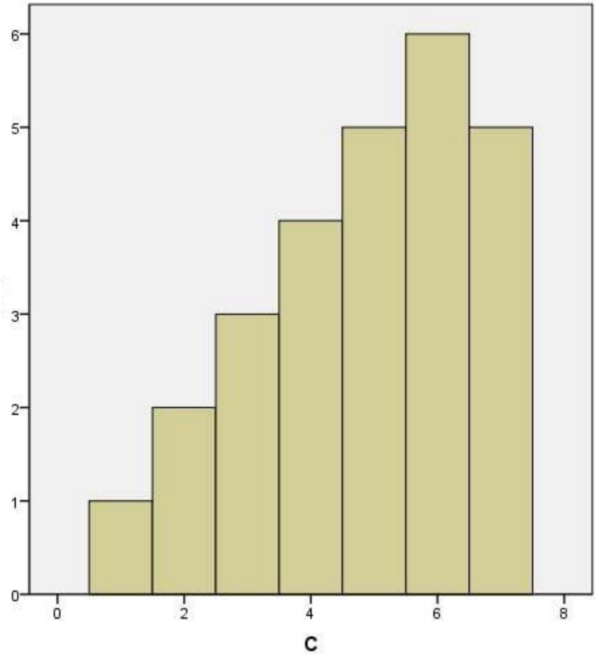
众数、中位数和平均数的关系



众数 = 中位数 = 平均数



众数 < 中位数 < 平均数



众数 > 中位数 > 平均数

算术平均数的求和稳定性

定理 设一批数据为X_1，X_2，…，X_n，$\bar{X}$为其算术平均数为,则

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

算术平均数

假设某射击运动员在某次射击训练中,分别打出了8环、9环和10环。环数和对应权重如下表:

环 数	8	9	10
权 重	0.2	0.65	0.15

则其加权算术平均数为 $\frac{8*0.2+9*0.65+10*0.15}{0.2+0.65+0.15}$=8.95

在实际应用中,权重的设置通常要求代表某种性质,如重要性、频繁度等等。在上例中, 权重代表了相应环数的频率:$\frac{20}{100}$, $\frac{65}{100}$, $\frac{15}{100}$。事实上,此时我们用算术平均数计算,其结果相同。

$$\frac{\sum_{i=1}^{20} 8+\sum_{i=1}^{65} 9+\sum_{i=1}^{15} 10}{100}=8.95$$

几何平均数

几何平均数是n个数据相乘的n次方根。在实际应用中,几何平均数主要用于描述平均比率(如平均增长率、平均速度等)。计算几何平均数通常要求每一个X_i非负,以确保n次方根有实际意义。

设一批数据为$X_1, X_2,\ldots,X_n$， ,则其几何平均数G_n为

$$G_n = \sqrt[n]{X_1 * X_2 * \ldots * X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

几何平均数

南京市某区的5年的经济增长率分别为8%,9%,10%,12%,和12%,则这5年的平均经济增长率计算如下:

$$\bar{X}_G = \sqrt[5]{1.08 * 1.09 * 1.10 * 1.12 * 1.12} \approx 1.102$$

$$1.102-1=0.102=10.2\%$$

我们可以这样理解增长率的几何平均数:设该区最初的经济总量为X_0,则第1年后的经济总量$X_1=X_0 *1.08$,第5年后的经济总量为$X_0 *1.08 * 1.09 * 1.10 * 1.12 * 1.12$,设平均增长率为I,则$X_0$ *(1.08*1.09*1.10*1.12*1.12)=$X_0 * (1+I)^5$,

$$I= \sqrt[5]{1.08 * 1.09 * 1.10 * 1.12 * 1.12}-1$$

对数(几何平均数) -> 算术平均数

调和平均数

调和平均数,也称为倒数平均数。它是对每个数据的倒数求平均,然后再求倒数得到的平均数。

设一批数据为$X_1, X_2,\ldots,X_n$， ,则其调和平均数H_n为

$$H_n = \frac{1}{\frac{\frac{1}{X_1}+\frac{1}{X_2}+\ldots+\frac{1}{X_n}}{n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}} \tag{9}$$

平方平均数

平方平均数是对每个数据的评分求平均进行开根号。

设一批数据为$X_1, X_2,\ldots,X_n$， ,则其调和平均数Q_n为

$$Q_n = \sqrt{\frac{X_1^{\ 2}+X_2^{\ 2}+\cdots+X_n^{\ 2}}{n}} \quad (10)$$

平均数不等式

定理

$$H_n \leq G_n \leq A_n \leq Q_n$$

离散趋势度量

- 集中趋势度量经常是我们产品设计的目标。
- 例如,我们设计的地铁交通系统能满足平均每小时1万人次的客流;我们设计的网站系统能满足平均每天10万访问量;电话交换系统只能容纳平均每小时10万次的呼叫量等等。
- 但是仅仅考虑数据的平均水平是不够的,我们还得考虑数据的离散程度，即系统能承受的波动。

离散趋势度量

- 研究数据的波动对于统计分析往往是必需的。在很多时候,我们考虑单独数据是没有意义的,只有将它与同分布的其它数据相比较才有意义。
- 例如,我们知道某班的考试成绩平均分为80分,一个人考了85分,比平均分高5分。然而他算优秀还是良好?这5分的差距大吗?这得取决于其它数据的分布。
- 如果我们知道了数据的集中趋势度量和离散趋势度量,就能更好的理解数据,做出正确判断。

全距

全距,又称极差,是样本观察值的最大值和最小值的差。全距是最简单的种离散趋势度量,通常用来反映一批数据总的离散趋势。但全距不考虑数据的分布情况。当数据中存在极值并不太关注数据分布时,全距是一种合适的离散趋势度量。

全距

例如,生产标准规格为100cm的钢条,甲乙两条生产线的钢条长度如下。

甲	101 98 100 102 99
乙	101 91 100 109 99

已知甲乙生产线的钢条平均长度都是100。如何比较那条生产线更稳定(产品长度均匀)?

甲的全距=102-98=4

乙的全距=109-91=18

内距

内四分位距,也称内距,内四分位距是用来描述中间50%的样本观察值的离散趋势度量。它是样本观察值的第三四分位数与第一四分位数的差。

$$内四分位距=Q_3-Q_1 \tag{11}$$

不受极值影响的度量值通常称为抵抗度量。虽然内距比全距更有意义,但它仍有以下两个缺点:

- 不能提供精确的数据分布信息。
- 不能用来进行精确的统计推断。

偏差平方和

为了更加全面考虑整体数据波动，我们首先引入数据的偏差平方和。

设一批数据为$X_1, X_2, \ldots, X_n$，则数据的偏差平方和为每个数据与平均数偏差平方的和：

$$d^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

算术平均数的偏差极小性

定理 设一批数据为X_1，X_2，…，X_n，$\bar{X}$为其算术平均数,则

$$\min_{x}\sum_{i=1}^{n}(X_i-X)^2=\bar{X}$$

方差与标准差

方差s^2为偏差平方和的平均数:

$$s^2 = \frac{d^2}{n} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

数据的标准差s为s^2的平方根:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}}$$

定义式与计算式

计算偏差平方和常用以下公式。

定理 设一批数据为$X_1,X_2,\ldots,X_n$，则其偏差平方和为d^2, 方差为s^2，则

$$d^2 = \sum_{i=1}^{n} {X_i}^2 - \frac{\sum_{i=1}^{n} {X_i}^2}{n}$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \bar{X}^2$$

开放练习

- 对数在数据预处理中的作用？
- 如何建立集中趋势度量和离散趋势度量之间的关系？