MAT220 第3讲：集中趋势与变异

Contents

在第1讲中，我们介绍了统计学的基本语言，包括数据、变量、总体、样本和抽样方法。在第2讲中，我们学习了如何用直方图、条形图、饼图、时间序列图和散点图来直观表示数据。在本讲中，我们开始进入描述统计的数值部分。我们的主要目标是学习如何用少数几个精心选择的统计量来概括一个数据集。我们将重点讨论两个主要概念：集中趋势，它描述数据的中间位置在哪里；以及变异，它描述数据分散到什么程度。

1 集中趋势

在本讲中，我们将处理定量数据。我们需要以一种有结构的方式讨论这些数据，因此首先要引入一些重要记号。在本讲使用的所有公式中，我们会用字母 x 表示数据值，并用下标给它们编号：

$$x_1,\ x_2,\ x_3,\ x_4,\ \ldots,\ x_n.$$

下标只是一个方便以后引用的标签。注意，数据的标签从 1 开始，一直数到 n。最后这个数 n（无论它具体是多少）表示数据集的总大小。当顺序很重要时，我们会把数据集从小到大写出来，也就是：

$$x_1 \le x_2 \le x_3 \le x_4 \le \cdots \le x_n.$$

1.1 集中趋势的含义

当我们收集了大量数据时，通常希望用一个有代表性的数来概括它。一种方法是描述数据的“中间”在哪里。这样，我们就可以说明数据点倾向于出现在什么位置。这正是集中趋势的思想。

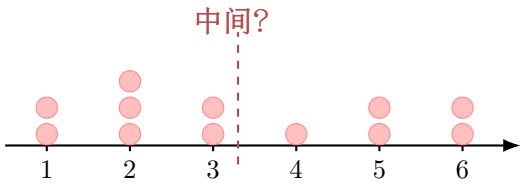
定义：集中趋势

集中趋势是指试图描述一个数据集的中间位置。

例如，考虑数据集

$$(1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6).$$

我们可以把每个观测值表示成一个小球，从而把这些数据画在数轴上：



在图中，数据的中间应该是这条虚线吗？还是应该在别的地方？集中趋势的任务就是找出一种有意义的方法来确定中间的位置。直觉上这很清楚，但数学问题并没有那么简单，因为“中间”这个词可以有许多种解释。因此，集中趋势有几种不同的度量方式。

1.2 求和记号

在后面要引入的公式中，我们需要对数据进行一些算术运算，因此也需要知道怎样高效地把这些运算写下来。为了说明这一点，假设我们正在处理一个大小为 $n = 1000$ 的数据集，并且想把所有数据值加起来。在这种情况下，我们不想反复手写：

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11} + x_{12} + \ldots$$

为了解决这个问题，我们使用求和记号。

求和记号

$$\sum_{i=1}^{n} x_i \quad \text{表示} \quad x_1 + x_2 + x_3 + \cdots + x_n.$$

字母 i 是**索引**。它告诉我们当前正在加哪一项。下面的数告诉我们从哪里开始，上面的数告诉我们在哪里停止。

从上面可以看出，求和记号只是写出一个很长的和的一种高效方式。需要注意的是，我们也可以用这个记号写出不只是简单相加 x_i 的情况。例如，我们可以写：

$$\sum_{i=1}^{n} 2x_i = 2x_1 + 2x_2 + 2x_3 + \cdots + 2x_n.$$

在这里，我们不是把所有单独的 x_i 值加起来，而是把所有项 $2x_i$ 加起来。

例子。 假设

$$(x_1, x_2, x_3) = (4, 10, 12).$$

那么

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 4 + 10 + 12 = 26,$$

$$\sum_{i=1}^{2} x_i = x_1 + x_2 = 4 + 10 = 14,$$

并且

$$\sum_{i=2}^{3} x_i^2 = x_2^2 + x_3^2 = 10^2 + 12^2 = 100 + 144 = 244.$$

练习

假设

$$(x_1, x_2, x_3, x_4, x_5) = (2, 4, 6, 8, 10).$$

计算：

(a) $\sum_{i=1}^{4} x_i$

(b) $\sum_{i=3}^{5} x_i$

(c) $\sum_{i=1}^{5} x_i$

(d) $\sum_{i=2}^{3} x_i$

(e) $\sum_{i=1}^{3} \frac{x_i}{2}$

解答

(a) $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = 2 + 4 + 6 + 8 = 20$。

(b) $\sum_{i=3}^{5} x_i = x_3 + x_4 + x_5 = 6 + 8 + 10 = 24$。

(c) $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 2 + 4 + 6 + 8 + 10 = 30$。

(d) $\sum_{i=2}^{3} x_i = x_2 + x_3 = 4 + 6 = 10$。

(e) $\sum_{i=1}^{3} \frac{x_i}{2} = \frac{x_1}{2} + \frac{x_2}{2} + \frac{x_3}{2} = \frac{2}{2} + \frac{4}{2} + \frac{6}{2} = 1 + 2 + 3 = 6$。

1.3 三种常见平均：均值、中位数和众数

在日常语言中，“平均”这个词是有歧义的。它可以指几种不同的统计量。现在我们将回顾三种最常见的平均：**均值**、**中位数**和**众数**。它们都是集中趋势的度量，但回答的问题略有不同。

统计量	基本思想	主要弱点
均值	数据的平衡中间位置。	对异常值敏感。
中位数	排序后的中间值。	不使用每个值的具体数值大小。
众数	出现最频繁的值。	可能不唯一，并且对数值数据来说未必有太多信息。

我们现在逐一回顾这些概念，从中位数开始。

1.3.1 中位数

中位数是把整个数据集分成两半的值。找到中位数后，至少一半的数据位于中位数或其以下，至少一半的数据位于中位数或其以上。

定义：中位数

中位数是有序数据集的中间值。

为了找到中位数，我们首先要注意 n 的值：如果它是奇数或偶数，我们就需要执行略有不同的操作。具体步骤是：

Step 1: 将所有值从小到大排序。
Step 2: 如果 n 是奇数，选取中间项。这是第 $\frac{n+1}{2}$ 项。
Step 3: 如果 n 是偶数，找出两个中间项，并计算它们正中间的数。

例如，考虑数据集：

$$(1, 2, 3, 4, 5).$$

这里有 5 个数据点，也就是说 $n = 5$。因此，为了找到中位数，我们给 n 加上 $+1$，然后除以二：$\dfrac{5+1}{2} = 3$。所以，我们只需要找到数据集中的第 3^{rd} 项，并读出它的值。在这个例子中，中位数是 3。

现在考虑数据集：

$$(2, 3, 4, 5, 6, 7).$$

这里有 6 个数，也就是说 $n = 6$，它是偶数。中位数的位置是 $\dfrac{6+1}{2} = 3.5$。这意味着中位数将是位于数据集第 3^{rd} 项和第 4^{th} 项正中间的数。观察可知，这两项分别是数据点 4 和 5。因此，中位数是这两个数的中间值，也就是 $\dfrac{4+5}{2} = 4.5$。

1.3.2 均值

均值就是我们通常所说的数据集的“平均”。对于样本

$$x_1, x_2, \ldots, x_n,$$

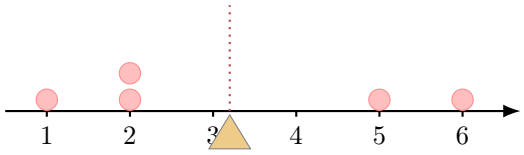
均值记为 $\bar{x}$，定义如下。

定义：样本均值

对于样本 $x_1, x_2, \ldots, x_n$，**样本均值**是

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

换句话说，均值 $\bar{x}$ 的计算方法是把数据集中的所有值加起来，然后除以值的个数。从几何上看，我们可以把均值想象成数据的平衡点。如果把每个数据点想象成放在数轴上的一个重物，那么均值就是整个系统达到平衡的位置。例如，对于数据集 $(1, 2, 2, 5, 6)$，我们可以把均值想象成中间的平衡点：



在这个例子中，数值上我们计算：

$$\bar{x} = \frac{1+2+2+5+6}{5} = \frac{16}{5} = 3.2.$$

均值的一个重要性质是，关于均值的正偏差和负偏差总会相互抵消。

关于均值的一个有用事实

对于任意数据集，

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

换句话说，从数据到均值的有符号距离总和总是零。

继续使用上面的数据集 $(1, 2, 2, 5, 6)$，我们可以手动验证这个事实。注意，均值 $\bar{x} = 3.2$ 左侧有三个值，即 $1, 2, 2$，而大于均值的值有两个，即 5 和 6。由于 $\bar{x}$ 左侧的值小于 3.2，所以 $x_i - \bar{x}$ 的值会是负数。按绝对值来看，实际距离是

$$(3.2 - 1) + (3.2 - 2) + (3.2 - 2) = 2.2 + 1.2 + 1.2 = 4.6,$$

而对于 $\bar{x}$ 右侧的数据点，我们有：

$$(5 - 3.2) + (6 - 3.2) = 1.8 + 2.8 = 4.6.$$

这些值说明，均值左侧的距离之和等于均值右侧的距离之和。这就是为什么把均值称为平衡中间位置是合理的。

另外，这个关于均值的事实也可以通过代数方法验证：

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x}) &= (x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})\\ &= x_1+x_2+\cdots+x_n-\bar{x}-\bar{x}-\cdots-\bar{x}\\ &= x_1+x_2+\cdots+x_n-n\cdot\bar{x}\\ &= \sum_{i=1}^{n}x_i-n\cdot\frac{\sum_{i=1}^{n}x_i}{n}\\ &= \sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i\\ &= 0.\end{aligned}$$

1.3.3 众数

简单地说，众数就是数据集中出现最频繁的值。

定义：众数

众数是数据集中出现次数最多的值。

如果有多个数据点以相同的最高频率出现，我们就说这个数据集没有唯一众数。

例子。 数据集

$$(1,1,1,3,4)$$

的众数是 1，因为 1 出现了三次。
数据集

$$(1,1,2,2,3)$$

没有唯一众数，因为 1 和 2 都出现了两次。

1.4 一个例题

考虑数据集

$$(1,1,2,2,3,4,5).$$

均值是

$$\bar{x}=\frac{1+1+2+2+3+4+5}{7}=\frac{18}{7}\approx 2.57.$$

有 7 个数据点，所以中间项是第

$$\frac{7+1}{2}=4$$

项，因此中位数是第 4 项。第 4 项是 2，所以中位数是 2。

值 1 和 2 都出现了两次，没有任何值出现得更多。因此，没有唯一众数。

练习

考虑数据集

$$(1, 1, 2, 2).$$

求：

(a) 均值，

(b) 中位数，

(c) 众数。

解答

(a) 均值是

$$\bar{x} = \frac{1+1+2+2}{4} = \frac{6}{4} = 1.5.$$

(b) 有 4 个数据点，所以两个中间值是第 2 项和第 3 项。它们是 1 和 2，所以中位数是

$$\frac{1+2}{2} = 1.5.$$

(c) 值 1 和 2 都出现了两次。因此，没有唯一众数。

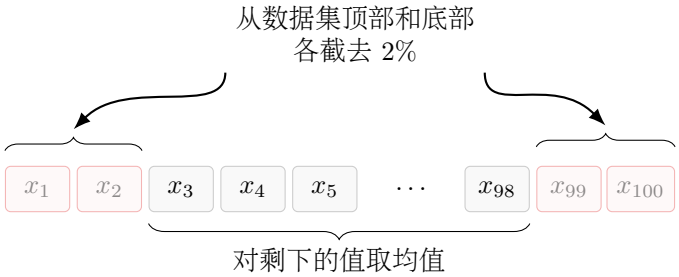
1.5 截尾均值

均值会使用数据集中的每一个值。这通常很有用，但也意味着均值会受到极端异常值的影响：如果有一个数据值远离一侧，那么均值为了保持平衡就必须作出很大的补偿。在某些情况下，均值的这个特征是可以接受的。然而，在其他情况下，我们可能希望有意去除极端异常值的影响。这就引出了截尾均值的概念。

定义：截尾均值

$x\%$ **截尾均值**是指从一个有序数据集中去掉最低的 $x\%$ 和最高的 $x\%$，然后计算剩余值的均值。

换句话说，我们通过切掉一个有序数据集底部的 $x\%$ 和顶部的 $x\%$，然后对剩余部分取均值，来得到 $x\%$ 截尾均值。例如，如果我们有一个大小为 100 的有序数据集，那么 2% 截尾均值会去掉两个最小值和两个最大值。然后，它会取剩下 96 个值的均值：



当我们试图平均一组人为判断时，截尾均值特别有用。例如，在花样滑冰中，一个评委组会给某

个个人表演打分。如果一名评委由于固有偏见或人为错误给出异常高分，而另一名评委给出异常低分，那么这些极端分数可能会不公平地影响均值。为了控制这种影响，可以用截尾均值来计算表演分数，从而减少这些极端判断的影响。

练习：截尾均值

考虑有序数据集

$$1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24.$$

这里 $n = 20$。

(a) 计算 5% 截尾均值。

(b) 计算 20% 截尾均值。

(c) 解释为什么这里的 50% 截尾均值没有意义。

解答

(a) 因为 20 的 5% 是 1，所以我们去掉最小值和最大值。剩余数据是

$$1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23.$$

这些值的和是 149，所以 5% 截尾均值是

$$\frac{149}{18} \approx 8.28.$$

(b) 因为 20 的 20% 是 4，所以我们去掉最低的四个值和最高的四个值。剩余数据是

$$4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11.$$

这些值的和是 84，所以 20% 截尾均值是

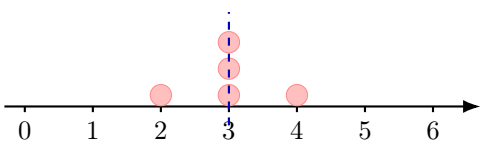
$$\frac{84}{12} = 7.$$

(c) 从底部截去 50%，再从顶部截去 50%，会去掉整个数据集。没有剩余值可以平均，因此 50% 截尾均值没有意义。

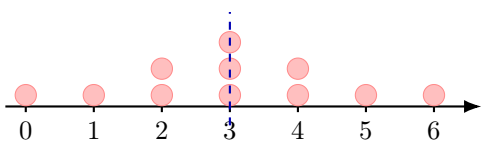
2 变异

集中趋势的度量可以让我们确定数据相对于某个中间位置倾向于在哪里。但是，集中趋势不能告诉我们数据分散到什么程度。例如，考虑下面画出的两个数据集：

$A = (2, 3, 3, 3, 4)$



$B = (0, 1, 2, 2, 3, 3, 3, 4, 4, 5, 6)$



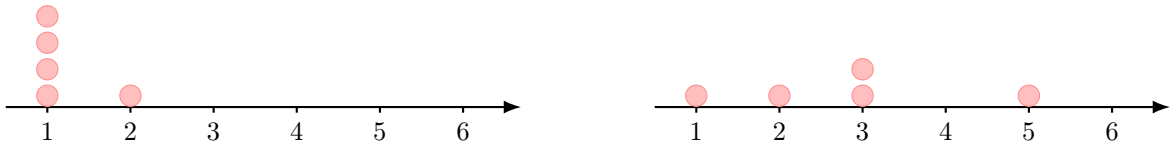
这两个数据集具有相同的均值（都等于 3），但数据集 B 显然比数据集 A 分散得多。仅仅描述数据的中心，永远无法捕捉分散程度这个特征。因此，我们需要引入一个新概念：变异。

定义：变异

变异描述数据分散到什么程度。

2.1 极差

考虑下面画出的两个数据集，$A = (1, 1, 1, 1, 2)$ 和 $B = (1, 2, 3, 3, 5)$。



从视觉上看，很明显 B 更分散，而 A 更集中。量化这一点的一种方法，是看数据集中最大值和最小值之间的距离：对于 A，整个数据集都包含在 1 和 2 之间；而对于 B，整个数据集从 1 扩展到 5。

这个简单观察引出了我们对分散程度的第一个定义：数据集的极差。

定义：极差

数据集的**极差**是

$$\text{极差} = \max(\text{数据}) - \min(\text{数据}).$$

例如，我们的数据集 $A = (1, 1, 1, 1, 2)$ 的极差是 $2 - 1 = 1$，而数据集 $B = (1, 2, 3, 3, 5)$ 的极差是 $5 - 1 = 4$。所以，极差确认了我们从视觉上看到的事实：B 比 A 更分散。

2.2 四分位数

“quarter”这个词表示某物的四分之一。与之不同，四分位数表示帮助我们把一个有序数据集分成四个部分的值。由于总共有四个部分，我们需要五条分界线来划分它们：中间三条，两端两条。

五个值

$$Q_0, \quad Q_1, \quad Q_2, \quad Q_3, \quad Q_4$$

分别标记最小值、第一四分位数、中位数、第三四分位数和最大值。

四分位数

对于一个有序数据集：

- Q_0 是最小值。
- Q_1 是数据下半部分的中位数。
- Q_2 是整个数据集的中位数。
- Q_3 是数据上半部分的中位数。
- Q_4 是最大值。

不同教材有时会使用略有不同的四分位数约定，尤其是当数据集有奇数个值时。在本课程中，我们将使用以下实用规则：先找到中位数，然后找到下半部分的中位数和上半部分的中位数。如果 n 是奇数，那么在求 Q_1 和 Q_3 时，我们不会把中位数包含在任何一半中。

2.3 四分位距

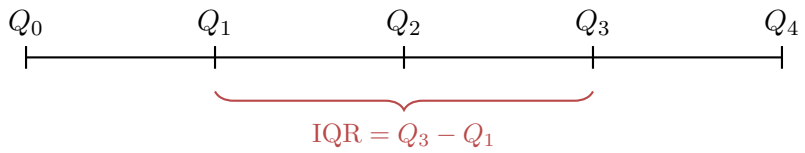
定义：四分位距

四分位距，缩写为 **IQR**，定义为

$$\mathrm{IQR} = Q_3 - Q_1.$$

它衡量数据中间一半的分散程度。

从视觉上看，四分位距衡量的是这段距离：



IQR 通常比极差更不容易受到异常值影响，因为它关注数据中间一半，并忽略较低的四分之一和较高的四分之一。

2.3.1 例题：四分位数和 IQR

考虑有序数据集：

$$(1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24).$$

为了计算四分位数，我们先找到中位数。因为总共有 20 个数据点，所以中位数位于第 10 项和第 11 项之间。这两项是 6 和 7，因此，中位数是 $Q_2 = \frac{6+7}{2} = 6.5$。为了找到 Q_1 的值，我们需要计算数据下半部分的中位数。写出来就是，我们要找下面这组数据的中间位置：

$$(1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6).$$

这个下半部分的中位数位于 4 和 5 之间，所以

$$Q_1 = \frac{4+5}{2} = 4.5.$$

类似地，数据的上半部分是：

$$(7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24).$$

这个上半部分的中位数位于 10 和 11 之间，所以 $Q_3 = \frac{10+11}{2} = 10.5$。

把这些放在一起，五数概括是：

$$Q_0 = 1, \qquad Q_1 = 4.5, \qquad Q_2 = 6.5, \qquad Q_3 = 10.5, \qquad Q_4 = 24,$$

四分位距可以计算为：

$$\mathrm{IQR} = Q_3 - Q_1 = 10.5 - 4.5 = 6.$$

2.4 方差

极差只使用两个数：最小值和最大值。这使它计算很快，但也意味着它忽略了数据集的大部分内容。极差还对异常值极其敏感。例如，考虑下面两个数据集：

$$A = (1, 1, 1, 2, 2, 2, 3, 6) \qquad B = (1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6)$$

两个数据集具有相同的极差：$6 - 1 = 5$。然而，从视觉上看，仍然可以感觉到数据集 A 比数据集 B 更不分散。在数据集 A 中，大部分数据集中在 $1, 2, 3$ 附近，只有一个异常值导致极差等于 5。相反，在数据集 B 中，数据分布得更加均匀，而且没有异常值。

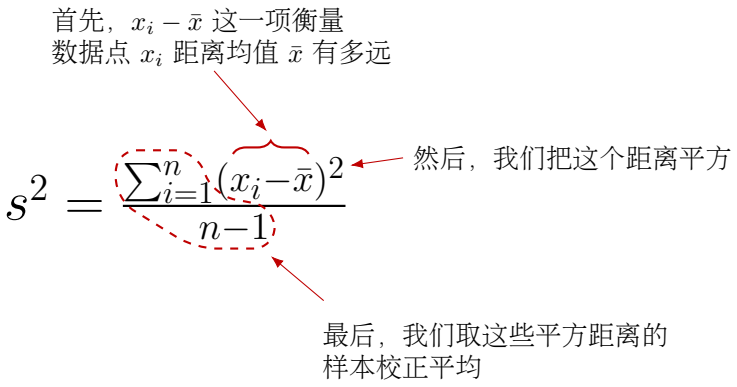
我们刚刚发现了数据集极差的一个根本缺陷：虽然它作为变异的初步度量很有用，但极差对异常值极其敏感，因此可能错过数据集的内部结构。为了解决这个问题，我们现在引入一种更细致的分散程度度量，称为方差。

定义：样本方差

对于样本 $x_1, x_2, \ldots, x_n$ 及其样本均值 $\bar{x}$，**样本方差**是

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

可以看到，这个公式比我们之前见过的公式复杂得多。因此，我们将一步一步拆解它，解释 s^2 实际上想表达什么。下面是这个公式的注释：



简而言之，方差 s^2 衡量的是每个数据值与均值之间的平方距离的样本校正平均：

1. 表达式 $x_i - \bar{x}$ 衡量点 x_i 离中心 $\bar{x}$ 有多远。它并不完全是距离，因为它可以取负值，而是称为偏差。
2. 表达式 $(x_i - \bar{x})^2$ 将所有这些偏差平方。
3. 接下来，求和记号 $\sum_{i=1}^{n}(x_i - \bar{x})^2$ 表示把所有这些平方偏差加起来，也就是

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2.$$

4. 最后，我们除以 $n - 1$，得到样本方差。

接下来，我们将解释为什么公式尤其是这个形式。

2.4.1 为什么要把偏差平方?

问为什么方差公式要把偏差平方是很自然的。为什么不直接把相对于均值的原始偏差加起来呢? 事实上，问题隐藏在我们前面见过的均值的重要性质中：

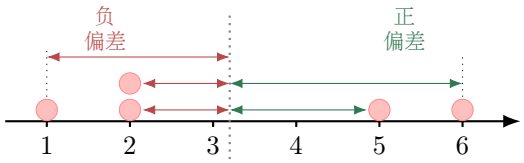
$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

这说明，所有原始偏差的和总是等于零。如果我们尝试使用未平方的原始偏差，那么总会得到：

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{0}{n-1} = 0,$$

无论一开始的数据是什么。例如，考虑数据集：

数据：$(1, 2, 2, 5, 6)$



它的均值是 $\bar{x} = 3.2$。 如果我们尝试把原始偏差加起来，就会得到：

$$(1 - 3.2) + (2 - 3.2) + (2 - 3.2) + (5 - 3.2) + (6 - 3.2) = 0.$$

平方通常可以解决这个问题。一旦把偏差平方，负偏差和正偏差都会得到非负的平方：

$$(-2)^2 = 4, \qquad 2^2 = 4.$$

这意味着总和被迫成为非负数，从而避免了上面提到的问题。[1]

2.4.2 为什么除以 $n - 1$?

通常，当我们取平均时，会把值加起来，然后除以值的个数。因此，很自然会问，为什么样本方差除以 $n - 1$ 而不是 n。

简短的回答是，s^2 实际上是一个**样本统计量**，而不是**总体参数**。通常，我们使用样本来估计更大总体的分散程度。当我们在公式中使用样本均值 $\bar{x}$ 时，已经从样本中估计了一个量，这会使平方偏差平均来说略微偏小。除以 $n - 1$ 可以校正这种影响。

贝塞尔校正

在样本方差公式中使用 $n - 1$ 而不是 n，称为**贝塞尔校正**。在本课程中，主要观点是实用性的：

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

是样本方差的标准公式。

本课程中我们不会证明这个校正。现在，你只需要知道应该使用哪个公式，以及它在测量什么。

[1] 另外，平方会给绝对值较大的偏差特别大的权重。

2.4.3 例题：方差

例子。 考虑数据集

$$(1, 1, 2, 3).$$

先计算均值：

$$\bar{x} = \frac{1+1+2+3}{4} = \frac{7}{4} = 1.75.$$

现在计算偏差和平方偏差：

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-0.75	0.5625
1	-0.75	0.5625
2	0.25	0.0625
3	1.25	1.5625

平方偏差之和是

$$0.5625 + 0.5625 + 0.0625 + 1.5625 = 2.75.$$

由于 $n = 4$，我们除以 $n - 1 = 3$：

$$s^2 = \frac{2.75}{3} \approx 0.917.$$

2.5 标准差

方差很有用，但它有一个不方便的特点：它以平方单位来测量。例如，如果数据以分钟为单位，那么方差就以平方分钟为单位。这个量可能很难直接和原始数据联系起来。我们可以通过简单地取方差的平方根来解决这个问题，这样结果就会与原始数据具有相同单位。这正是标准差的定义。

定义：样本标准差

样本标准差是

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

标准差与数据具有相同单位。因此，它更容易被解释为到均值的距离。

在上一个例子中，我们求得

$$s^2 \approx 0.917.$$

因此，

$$s = \sqrt{0.917} \approx 0.958.$$

所以标准差大约是 0.958。

2.6 变异系数

有时，我们想比较尺度差异很大的数据集的分散程度。例如，如果均值是 4，标准差 2 很大；但如果均值是 2000，标准差 2 就很小。为了让分散程度相对于数据集的典型大小，我们可以使用变异系数。

定义：变异系数

对于正的比率尺度数据，变异系数是一种无单位的相对分散程度度量。它定义为比值:

$$\mathrm{CV} = \frac{s}{\bar{x}}.$$

更大的变异系数意味着相对于数值的典型大小而言有更多变异。

当均值为正且不接近零时，这个统计量最有用。如果均值非常接近零，那么除以均值可能会使变异系数具有误导性或不稳定。

2.7 样本统计量与总体参数

正如第1讲所讨论的，推断统计经常使用样本统计量来估计总体参数。因此，我们对样本和总体使用不同记号。

数量	样本统计量	总体参数
均值	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
方差	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
标准差	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

这里，n 是样本大小，而 N 是总体大小。注意，对于样本方差和样本标准差，我们使用了贝塞尔校正。但是对于总体方差或总体标准差，我们不需要这样做，因为我们不再是在进行估计。

3 箱线图

3.1 五数概括

箱线图是五数概括的一种视觉表示。

五数概括

数据集的**五数概括**是：

$$\text{最小值}, \quad Q_1, \quad \text{中位数}, \quad Q_3, \quad \text{最大值}.$$

等价地，

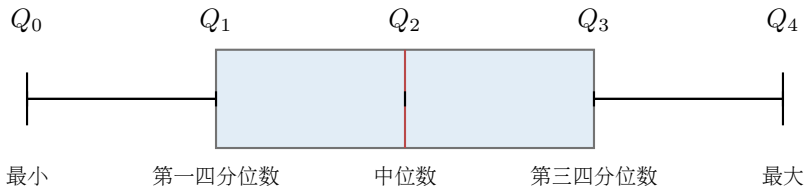
$$Q_0, \quad Q_1, \quad Q_2, \quad Q_3, \quad Q_4.$$

箱线图很有用，因为它同时显示集中趋势和变异。中位数显示数据的中间位置，箱体显示数据的中间一半，须显示数据的外侧部分。

3.2 箱线图的结构

在本讲使用的简单箱线图中：

- 左须从最小值 Q_0 开始，
- 箱体左边是 Q_1，
- 箱体内的线是中位数 Q_2，
- 箱体右边是 Q_3，
- 右须在最大值 Q_4 处结束。



箱体从 Q_1 延伸到 Q_3。须从 Q_0 延伸到 Q_1，并从 Q_3 延伸到 Q_4。

定义：箱线图

在本讲使用的简单版本中，**箱线图**是一种用五数概括来显示数据集的图形：最小值、Q_1、中位数、Q_3 和最大值。

3.3 构造箱线图

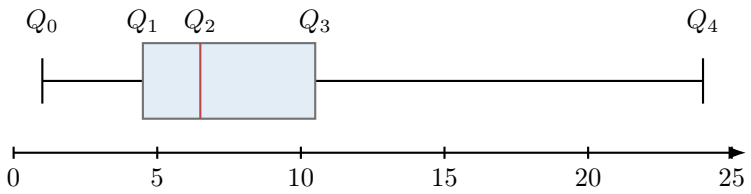
考虑数据集：

$$1,\ 1,\ 2,\ 3,\ 4,\ 5,\ 5,\ 6,\ 6,\ 6,\ 7,\ 7,\ 8,\ 9,\ 10,\ 11,\ 16,\ 20,\ 23,\ 24.$$

我们已经确定了五数概括：

$$Q_0 = 1, \qquad Q_1 = 4.5, \qquad Q_2 = 6.5, \qquad Q_3 = 10.5, \qquad Q_4 = 24.$$

由此得到的箱线图是：



注意，右须比左须长得多。这说明数据的上端比下端更加分散。