在使用均数、中位数以及众数衡量集中趋势时，用一个数来表示所有的数值，这个过程中损失了很多信息。我们不知道集合中的数字是接近该集中趋势，还是远离该集中趋势.

例如，有两组数据（2,2,3,3）和（0,0,5,5）。它们的均值都是2.5.但是第一组数据接近均值，第二组数据远离均值。也就是说，均值虽然用来衡量集中趋势，但不能很好地代表所有数字，当数据远离均值时，该如何衡量呢?

衡量的方法是方差（variance）。方差的符号是σ^2

> 方差是离中(dispersion)趋势的中的一个衡量指标

1. 总体方差

总体方差的公式：

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

> 方差公式的含义：求每个样本与均值的差的平方，然后相加后再除以数据个数。其中，样本与均值的差的平方也就是样本值与均值之间距离的绝对值的平方。

我们再来看上面两组数据的方差。

i	x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
1	2	2.5	-.5	0.25
2	2	2.5	-.5	0.25
3	3	2.5	.5	0.25
4	3	2.5	.5	0.25

第一组数据的方差为0.25

i	x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
1	0	2.5	-2.5	6.25

2	0	2.5	-2.5	6.25
3	5	2.5	2.5	6.25
4	5	2.5	2.5	6.25

第二组数据的方差为6.25

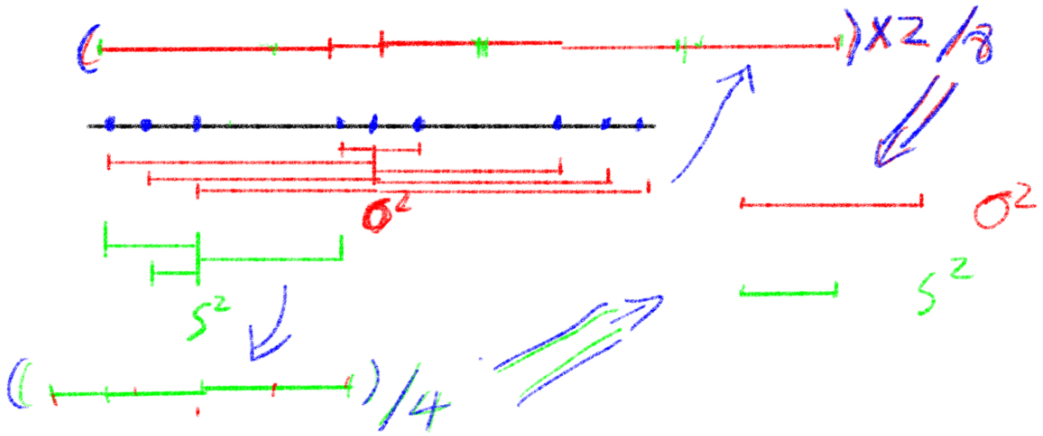
第二组数据的方差比第一组数据的方差大的多。这从直观上告诉我们，第二组数据平均离均值比另一组远得多（the 2td set are, on average, much further away from the mean than the numbers in the 1td set）

2. 样本方差

当总体不可测量时，我们可以用样本来估计总体。样本方差的公式：

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

这里是用n-1作为分母，为什么？因为抽样时，如果采样点分布的不是很好，样本与均值的距离之和通常是小于总体的数据与均值之间的距离之和的。如下图：



所以，为了无偏差估计，分母采用n-1，可以使方差大一点。

3. 方差公式的简化

方差公式的简化：

$$
\begin{aligned}
\sigma^2 &= \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \\
&= \frac{\sum_{i=1}^{N}(x_i^2-2\mu x_i+\mu^2)}{N} \\
&= \frac{\sum_{i=1}^{N}x_i^2-2\mu\sum_{i=1}^{N}x_i+\mu^2\sum_{i=1}^{N}1}{N} \\
&= \frac{\sum_{i=1}^{N}x_i^2}{N}-\frac{2\mu\sum_{i=1}^{N}x_i}{N}+\frac{\mu^2\sum_{i=1}^{N}1}{N} \\
&= \frac{\sum_{i=1}^{N}x_i^2}{N}-2\mu\mu+\mu^2 \\
&= \frac{\sum_{i=1}^{N}x_i^2}{N}-\mu^2
\end{aligned}
$$

在计算机中实现时，可以用下面的公式，该公式不需要计算均值

$$
\begin{aligned}
\sigma^2 &= \frac{\sum_{i=1}^{N}x_i^2}{N}-\mu^2 \\
&= \frac{\sum_{i=1}^{N}x_i^2}{N}-\frac{(\sum_{i=1}^{N}x_i)^2}{N^2}
\end{aligned}
$$