为什么样本方差（sample variance）的分母是 n-1？

先把问题完整的描述下。

如果已知随机变量 X 的期望为 μ，那么可以如下计算方差 σ^2：

$$\sigma^2 = E[(X-\mu)^2]$$

上面的式子需要知道 X 的具体分布是什么（在现实应用中往往不知道准确分布），计算起来也比较复杂。

所以实践中常常采样之后，用下面这个 S^2 来近似 σ^2：

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

其实现实中，往往连 X 的期望 μ 也不清楚，只知道样本的均值：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$

那么可以这么来计算 S^2：

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

那这里就有两个问题了：

- 为什么可以用 S^2 来近似 σ^2？
- 为什么使用 $\overline{X}$ 替代 μ 之后，分母是 $\dfrac{1}{n-1}$？

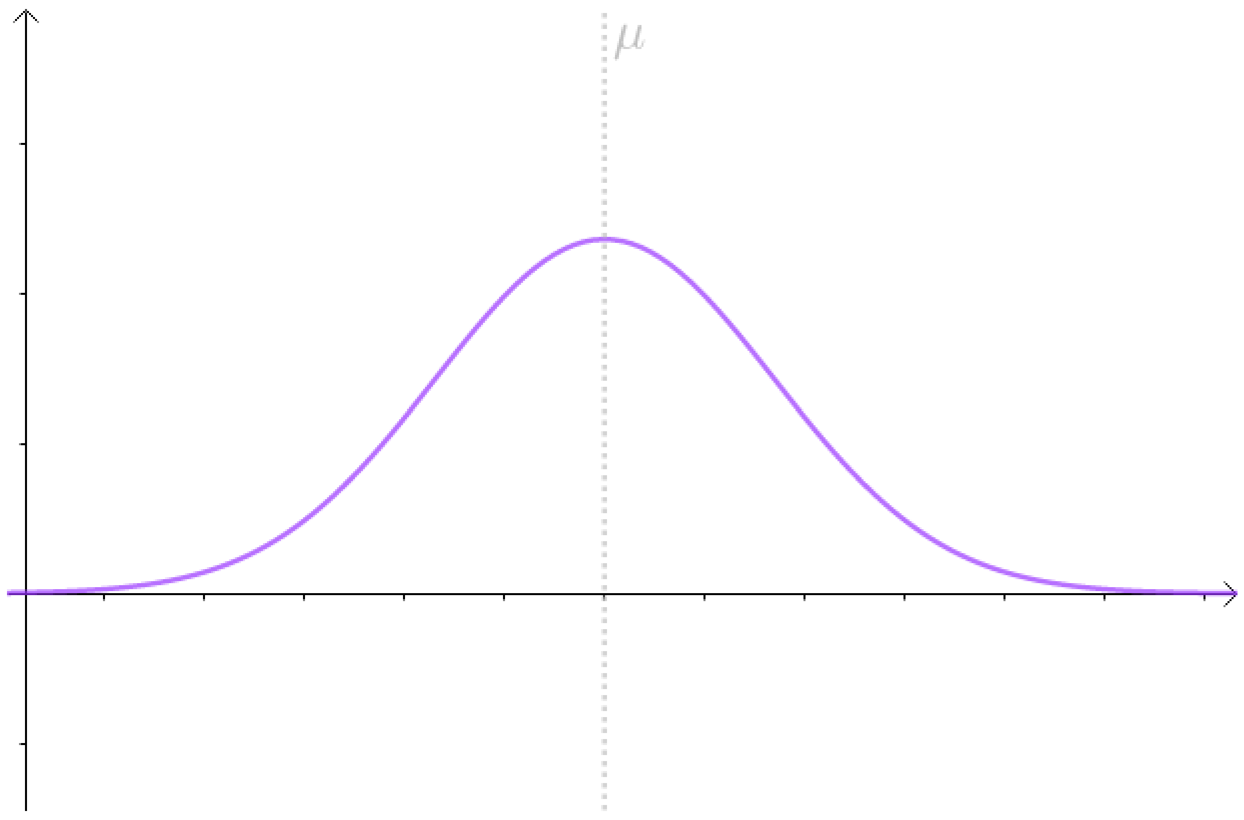
我们来仔细分析下细节，就可以弄清楚这两个问题。

1 为什么可以用 S^2 来近似 σ^2？

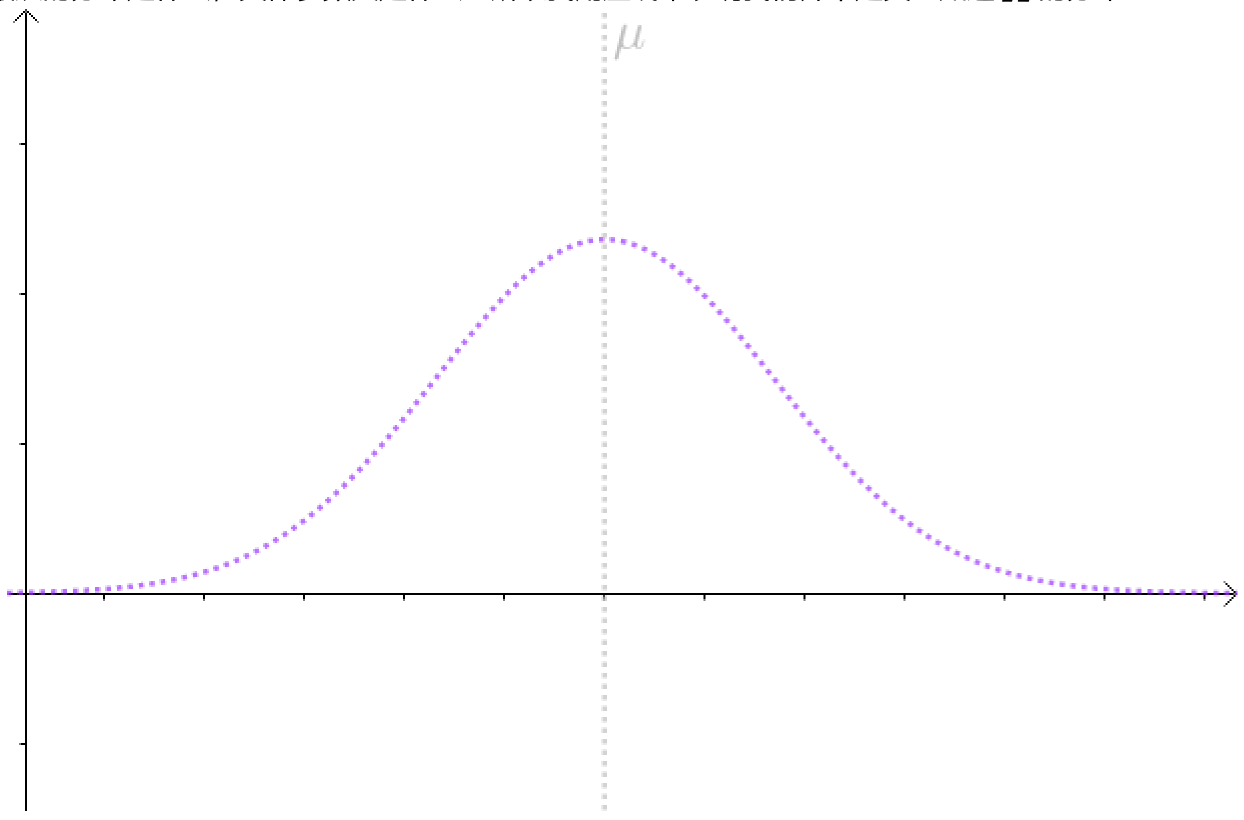
举个例子，假设 X 服从这么一个正态分布：

$$X \sim N(145, 1.4^2)$$

即，$\mu = 145, \sigma^2 = 1.4^2 = 1.96$，图形如下：



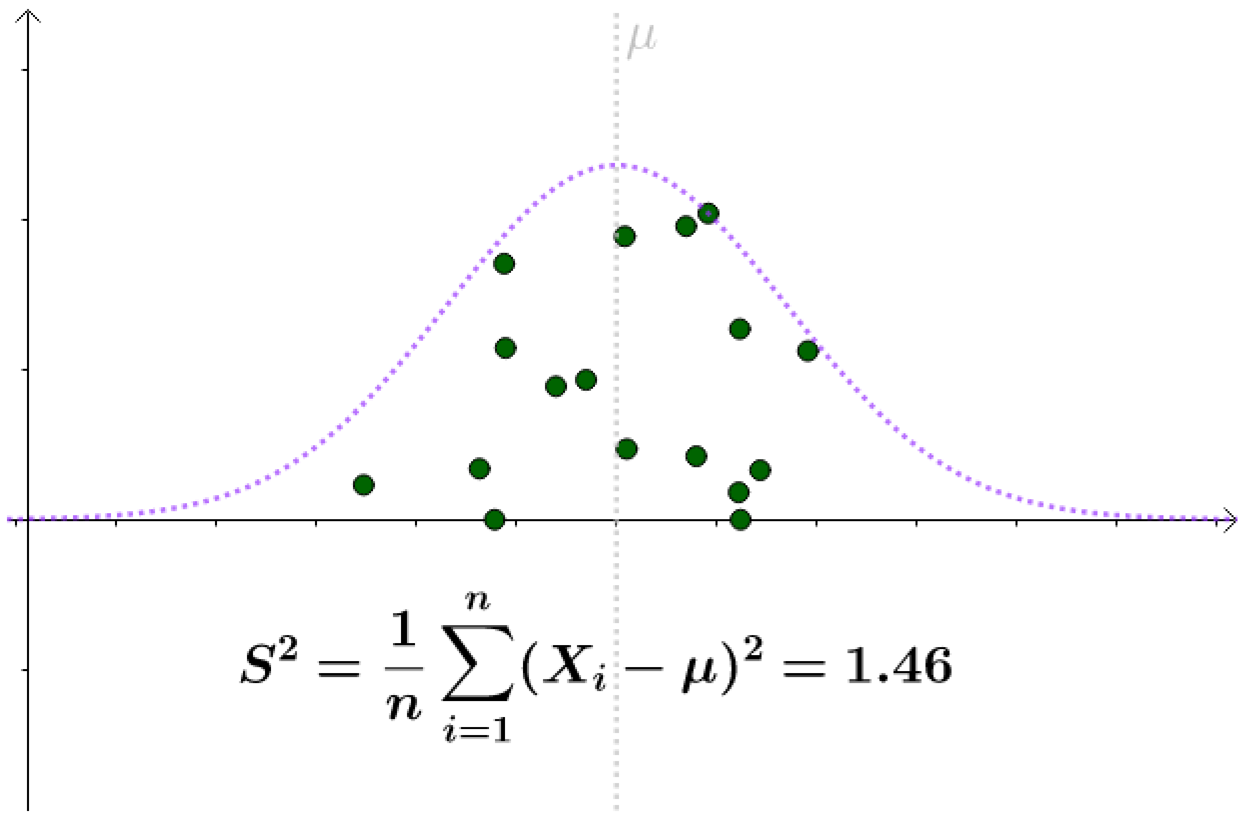
当然，现实中往往并不清楚 X 服从的分布是什么，具体参数又是什么？所以我用虚线来表明我们并不是真正知道 X 的分布：



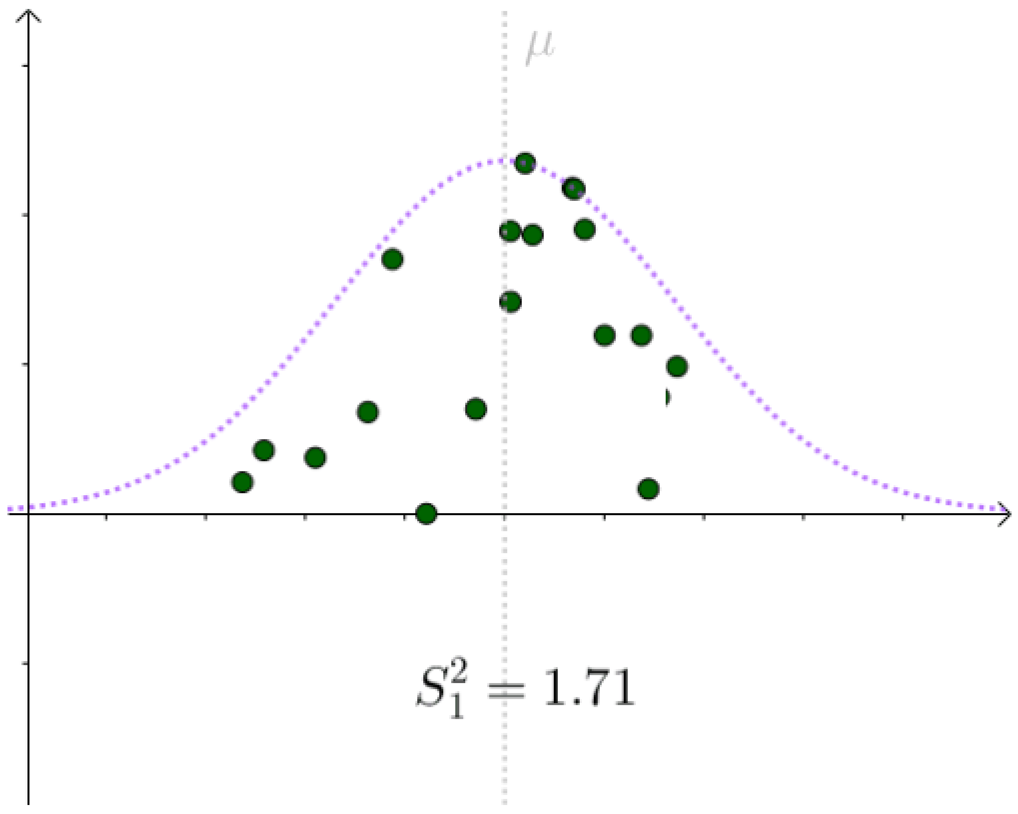
很幸运的，我们知道 $\mu = 145$，因此对 X 采样，并通过：

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

来估计 σ^2。某次采样计算出来的 S^2：



看起来来比 $\sigma^2 = 1.96$ 要小。采样具有随机性，我们多采样几次，S^2 会围绕 σ^2 上下波动：

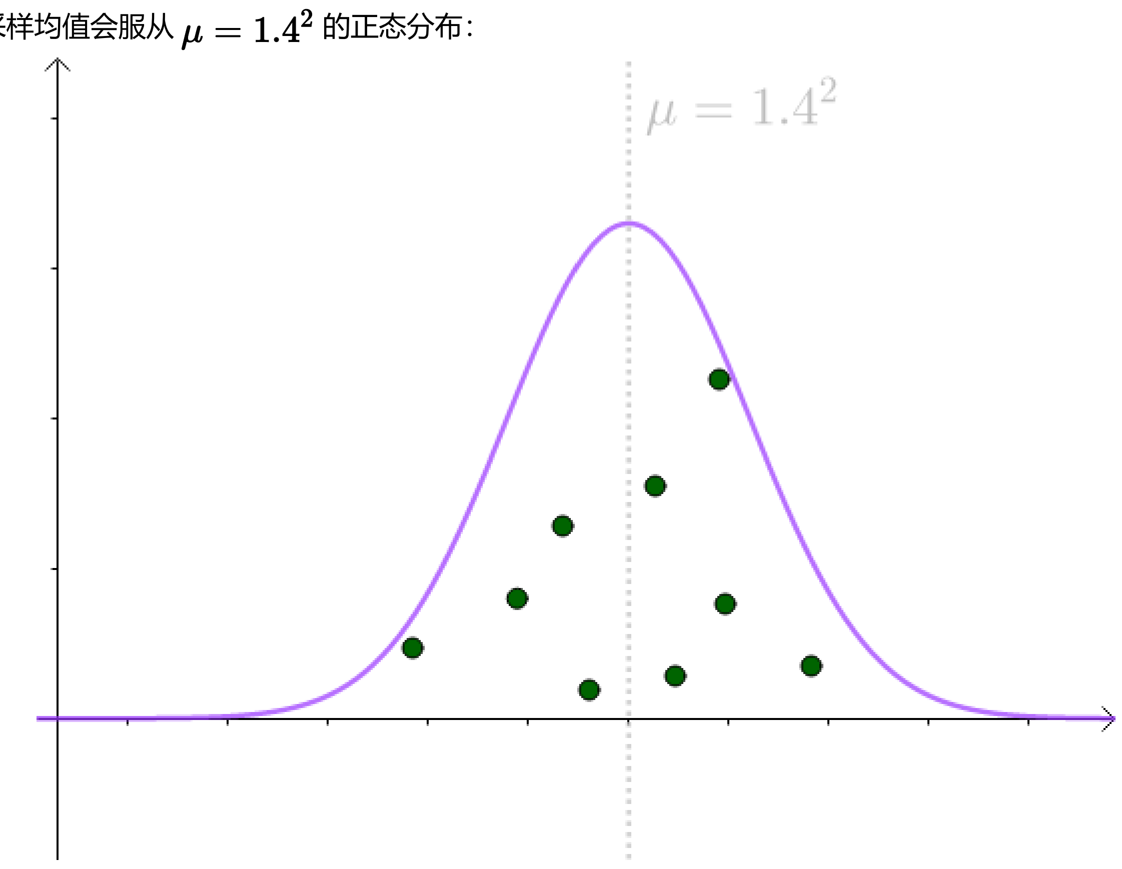


用 S^2 作为 σ^2 的一个估计量，算是可以接受的选择。

很容易算出：

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2] = \sigma^2$$

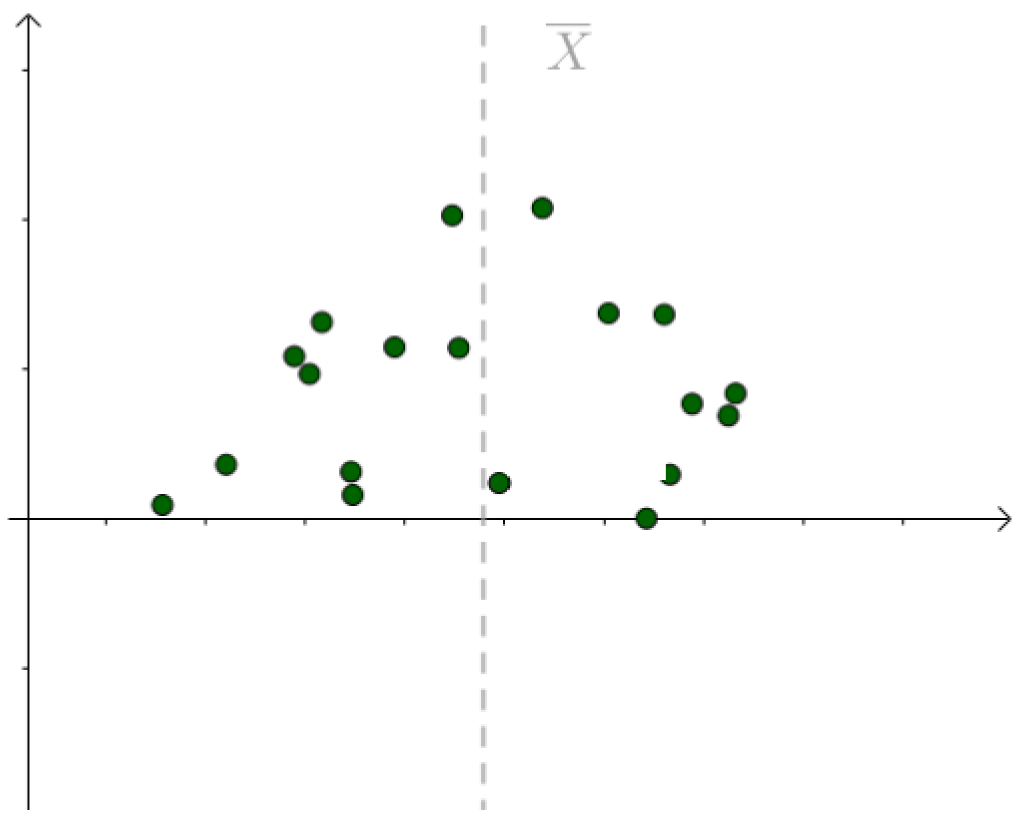
因此，根据中心极限定理，S^2 的采样均值会服从 $\mu = 1.4^2$ 的正态分布：



这也就是所谓的无偏估计量。从这个分布来看，选择 S^2 作为估计量确实可以接受。

2 为什么使用 $\overline{X}$ 替代 μ 之后，分母是 $\frac{1}{n-1}$？

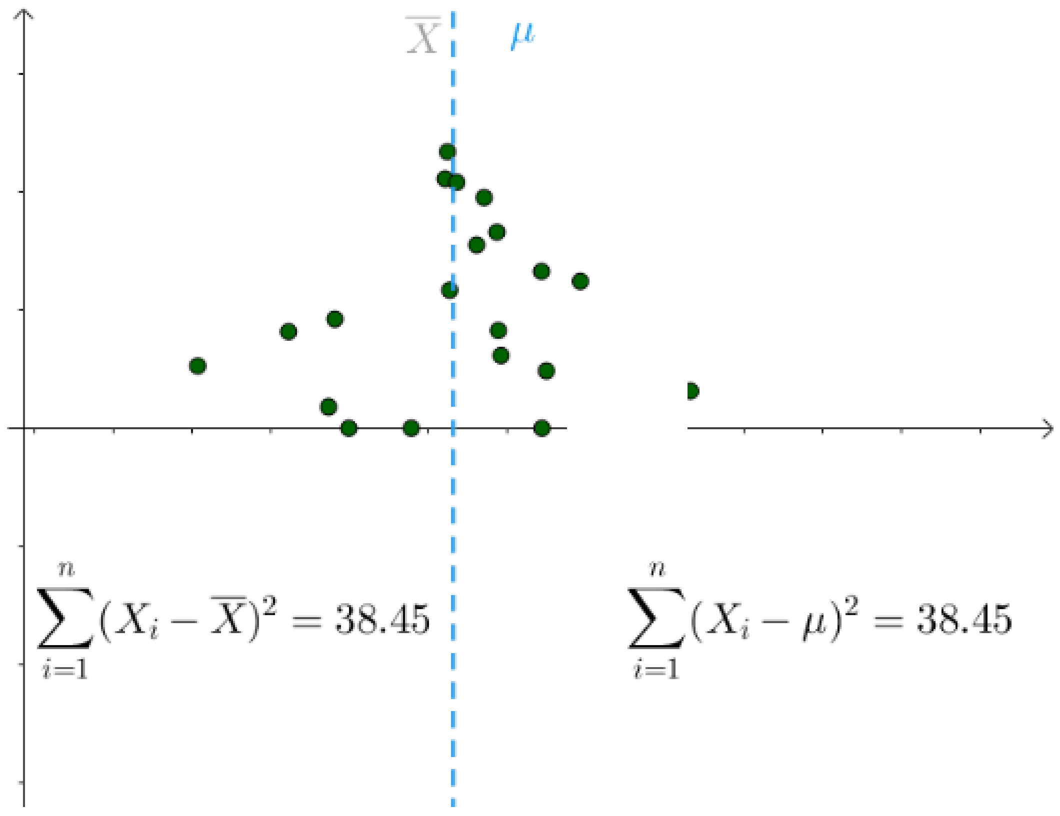
更多的情况，我们不知道 μ 是多少的，只能计算出 $\overline{X}$。不同的采样对应不同的 $\overline{X}$：



对于某次采样而言，当 $\mu = \overline{X}$ 时，下式取得最小值：

$$\sum_{i=1}^{n}(X_i-\mu)^2$$

我们也是比较容易从图像中观察出这一点，只要 μ 偏离 $\overline{X}$，该值就会增大：



所以可知：

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 \leq \sum_{i=1}^{n}(X_i-\mu)^2$$

可推出：

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \leq \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

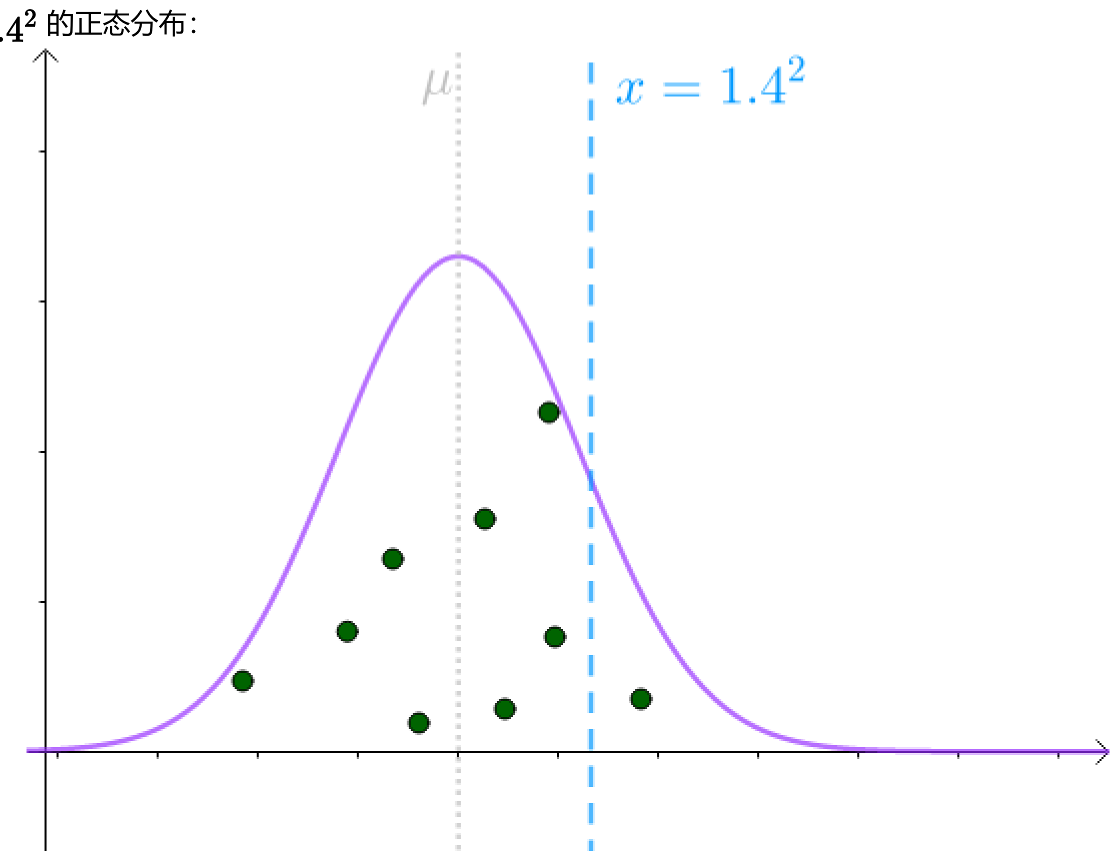
进而推出：

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2] \leq E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2] = \sigma^2$$

如果用下面这个式子来估计：

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

那么 S^2 采样均值会服从一个偏离 1.4^2 的正态分布：



可见，此分布倾向于低估 σ^2。

具体小了多少，我们可以来算下：

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bigg((X_i-\mu)-(\overline{X}-\mu)\bigg)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bigg((X_i-\mu)^2-2(\overline{X}-\mu)(X_i-\mu)+(\overline{X}-\mu)^2\bigg)\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\frac{1}{n}(\overline{X}-\mu)^2\sum_{i=1}^{n}1\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\frac{1}{n}(\overline{X}-\mu)^2\cdot n\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+(\overline{X}-\mu)^2\right]
\end{aligned}$$

其中：

$$\overline{X}-\mu = \frac{1}{n}\sum_{i=1}^{n}X_i-\mu = \frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu \ = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu).$$

所以我们接着算下去：

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) + (\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \frac{2}{n}(\overline{X}-\mu)\cdot n\cdot(\overline{X}-\mu) + (\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - 2(\overline{X}-\mu)^2 + (\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] - \mathrm{E}\left[(\overline{X}-\mu)^2\right] \\
&= \sigma^2 - \mathrm{E}\Big[(\overline{X}-\mu)^2\Big]
\end{aligned}$$

其中:

$$\mathrm{E}\left[(\overline{X}-\mu)^2\right] = \frac{1}{n}\sigma^2.$$

所以:

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2] = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

也就是说，低估了 $\frac{1}{n}\sigma^2$ ，进行一下调整:

$$\frac{n}{n-1}E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2] = E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2] = \sigma^2$$

因此使用下面这个式子进行估计，得到的就是无偏估计:

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