# 均方误差

虽然每一个观察函数都是一个估计量，但并不是每个估计量都是好的。好的 $g(\theta)$ 的估计量是一个观察数据的函数 $T$，使得 $T$ “接近”被估计的 $g(\theta)$。

那么怎么去讨论接不接近呢, 我们需要一个距离衡量. 如果单纯的使用距离 $\|T-g(\theta)\|$ 并不是一个令人满意的衡量标准，原因有两个：

- 这个度量依赖于未知的值 $\theta$, 我们都不知道这个"真"参数。何从谈及"距离"
- 这个度量是一个随机变量的，在实验进行(根据真实数据计算)之前是无法得知的。

我们先来解决第二个困难，也就是说先忽略第一个, 因此我们考虑在假设知道 $\theta$ 为真值时，距离 $\|T-g(\theta)\|$ 的分布。

最佳的情况是当该分布在 0 处退化，即如果 $\theta$ 是真值，那么 $\|T-g(\theta)\|$ 概率 1 等于 0。这将意味着我们不会产生任何估计误差；估计值 $T(x)$ 将绝对确定等于估计量。

这显然是一个完美的估计量, 不幸的是，在实际中这是不可能的，因此我们必须接受可能的（平均）误差, 但希望其最小, 也就是说找到最小平均误差。实践地说, 我们寻找一个估计量，使得在真值为 $\theta$ 的情况下，其分布尽可能集中在 $g(\theta)$ 周围，或者等价地，使 $\|T-g(\theta)\|$ 的分布尽可能集中在 0 的邻域内。

## 例 均匀分布

设 $X_1,\dots,X_n$ 为相互独立的 $\mathrm{U}[0,\theta]$ 分布的随机变量。观察值为向量 $X=(X_1,\dots,X_n)$，我们希望估计未知的 $\theta$。由于 $\mathrm{E}_\theta X_i=\frac{1}{2}\theta$，那么我们利用直观上"期望"的意义: "平均"来对样本进行操作. 也就是说用样本均值 $\bar{X}$ 来估计 $\frac{1}{2}\theta$ 是直观而且合理的，或者说通过 $2\bar{X}$ 来估计 $\theta$；毕竟，根据大数定律，样本均值以概率收敛到 $\mathrm{E}_\theta X_i=\frac{1}{2}\theta$。

假设 $n=10$，数据的取值为：$3.03, 2.70, 7.00, 1.59, 5.04, 5.92, 9.82, 1.11, 4.26, 6.96$，因此 $2\bar{x}=9.49$ 。

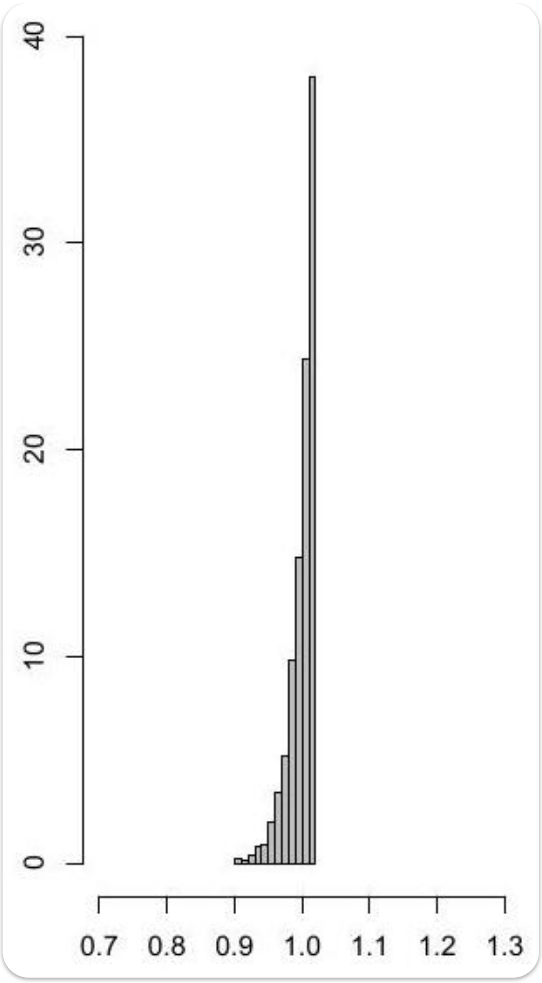
这个估计值显然太小了。仔细看观察一下的话会发现，其中一个观察值是 $9.82$，因此 $\theta\geq 9.82$ 是必须的(根据均匀分布的定义)。

我们能否想到一个更好的估计量呢？从上面的教训中学习的话, 我们可以通过取观察值的最大值 $X_{(n)}$ 来避免上述问题。然而，我们所取最大值也肯定小于(等于)真实值，因为所有观察值 $x_i$ 都位于区间 $[0,\theta]$ 之内。

一个浅显的解决方案是添加一个小的修正。例如，可以取 $(n+2)/(n+1)X_{(n)}$ 作为估计量, 这看上去合理多了。

因此我们现在手头有几个候选估计量。那么，哪个估计量是最好的呢？为了深入了解这个问题，我们进行了如下模拟实验。我们选择了 $n=50$ 并从 $[0,1]$ 的均匀分布中模拟了 $1000$ 个独立样本，每个样本包含 $50$ 个观察值。对于每个样本，我们分别计算了估计量 $2\bar{X}$ 和 $(n+2)/(n+1)X_{(n)}$。下图显示了两个集合中各 $1000$ 个对参数 $\theta$ 的估计值的直方图。左图使用估计量 $(n+2)/(n+1)X_{(n)}$，右图使用 $2\bar{X}$。

这些直方图可以视为估计量密度的近似。左边的密度比右边的更集中在真实值 $\theta = 1$ 附近。因此，我们更倾向于使用估计量 $(n+2)/(n+1)X_{(n)}$：它“平均而言”更接近真实值。（直方图的形式差异是显著的：左边的直方图类似于（反向）指数分布密度，而右边的类似于正态分布密度。可以轻松地从理论上解释这一点。想想为什么？）(答案: 可以计算两个统计量的密度, 一个是算的均值, 分布是指数分布, 另一个次序统计量最大值是正态分布)



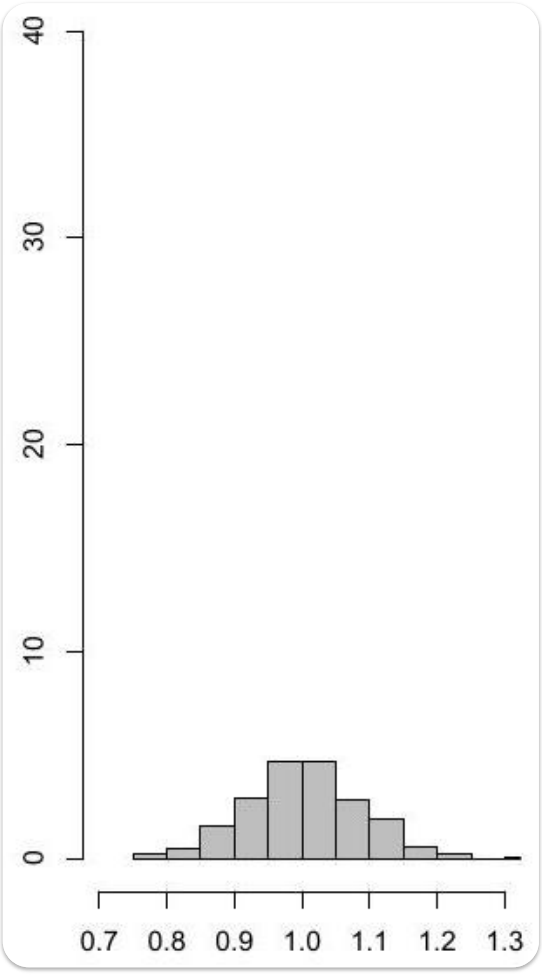




图: 基于 $n = 50$ 的观测值，估计量 $(n + 2)/(n + 1)X_{(n)}$ 和 $2\bar{X}$ 对均匀分布参数 $1$ 进行 $1000$ 次实现的直方图。

需要注意的是，估计量 $(n+2)/(n+1)X_{(n)}$ 并不总是在 1000 个样本中的每一个样本上给出最好的估计值。下图中可以看到这一点，其中竖直轴表示差值 $\left|(n+2)/(n+1)x_{(n)}-1\right|-|2\bar{x}-1|$。

可以看到的是通常情况下，这个差值是负的，但有时它是正的，也就是说此时估计量 $2\bar{X}$ 给出的值更接近真实值 $\theta=1$。由于在实践中我们不知道真实值，因此无法选择“最好的两个世界中的一个”。我们必须使用在平均情况下表现最好的估计量。

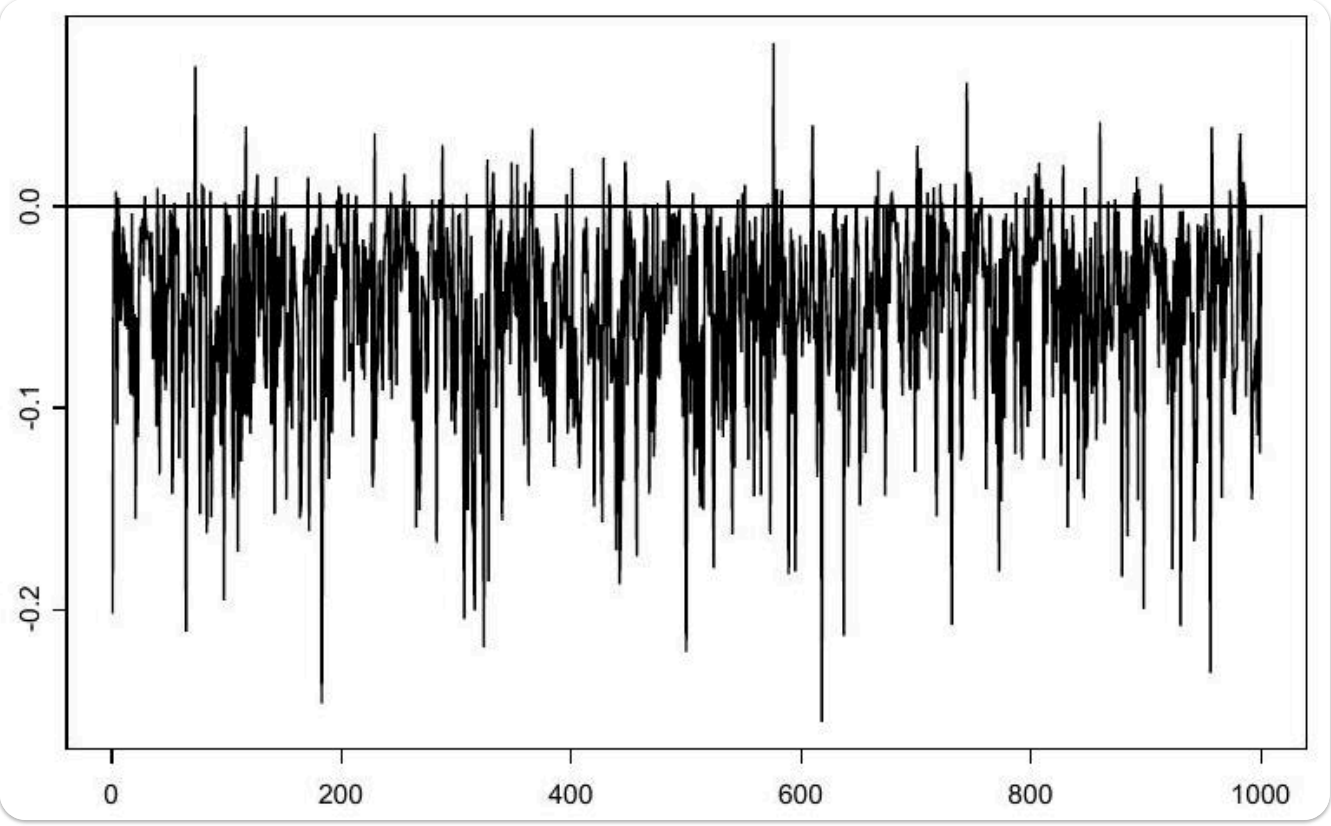



图 : 估计量 $(n+2)/(n+1)x_{(n)}$ 和 $2\bar{x}$ 与估计值 1 的绝对距离之差。

我们的模拟实验仅表明，当 $\theta$ 的真实值等于 1 时，$(n+2)/(n+1)X_{(n)}$ 是更好的估计量。要确定当 $\theta$ 具有不同值时哪个估计量更好，我们需要为每个 $\theta$ 从 $[0,\theta]$ 的均匀分布中模拟样本并重复进行模拟实验。

当然，这不是我们想做的事情，工作量实在太大, 所以我们需要数学工具研究估计问题的原因之一。另一个原因是我们不仅想对估计量对进行排序, 给出这比那好的相对评价，还希望找到总体上理论上最好的估计量。

由于概率分布是一个复杂的对象，比较我们刚刚提到的距离在 0 附近的“集中度”并没有明确的定义。因此，将集中度表达为一个数字是有用的，这样我们只需要比较数字即可。有许多方法可以做到这一点。一个在数学上相对简单的集中度衡量标准是均方误差或均方偏差。

### 均方误差

估计量 $T$ 对于值 $g(\theta)$ 的均方误差或 MSE 为

$$\mathrm{MSE}(\theta;T)=\mathrm{E}_\theta\|T-g(\theta)\|^2$$

定义中的下标 $\theta$ 是关键的：均方误差是 $T$ 在 $\theta$ 为真值时偏离 $g(\theta)$ 的期望平方误差。

我们将均方误差视为给定统计量 $T$ 的函数 $\theta \mapsto \mathrm{MSE}(\theta; T)$。更完整的表示法应为 $\mathrm{MSE}(\theta; T, g)$，但由于在问题的上下文中 $g$ 是固定的，我们省略了 $g$。

通过算期望和定义 $MSE$ 我们对第二个困难进行了回答, 那么还剩下第一个困难——衡量依赖于 $\theta$ ——尚未解决；均方误差是 $\theta$ 的函数。原则上，如果 $\mathrm{MSE}(\theta; T)$ 在 $\theta$ 的“真值”处尽可能小即可。由于我们不知道这个值，我们试图让均方误差在所有 $\theta$ 值上都保持（相对）较小。

**约定**

我们倾向于选择在所有参数值 $\theta$ 上均具有较小均方误差 (MSE) 的估计量。

对于两个估计量 $T_1$ 和 $T_2$，如果

$$\mathrm{E}_\theta \|T_1 - g(\theta)\|^2 \leq \mathrm{E}_\theta \|T_2 - g(\theta)\|^2 \quad \text{对所有 } \theta \in \Theta$$

且对至少一个 $\theta$ 存在严格不等式，则我们更倾向于选择 $T_1$。此时，估计量 $T_2$ 被称为不可接受的（inadmissible）。

然而实际上可能存在这样的情况，对于某些 $\theta$，严格不等式可能成立，而对其他 $\theta$，可能存在反向的不等式。这时我们无法直接判断该选择哪个估计量。由于 $\theta$ 的真实值 $\theta_0$ 是未知的，我们也无法确定 $\mathrm{MSE}(\theta_0; T_1)$ 和 $\mathrm{MSE}(\theta_0; T_2)$ 中哪个更小。

之后我们会讨论估计量的最优性准则以及如何找到最优估计量。现在我们先讨论几种寻找直观上合理的估计量的方法，并比较它们的均方误差。

**实际上估计量 $T$ 的均方误差可以分解为两项：**

$$\mathrm{MSE}(\theta; T) = \mathrm{var}_\theta T + \left(\mathrm{E}_\theta T - g(\theta)\right)^2$$

（可自行验证）

这个分解中的两项都是非负的。因此，均方误差只有在两项都较小时才会较小。如果第二项为 $0$ ，则估计量被称为无偏的, 叫无偏估计量。