无偏估计量

> **无偏估计量**
>
> 估计量 T 被称为估计 $g(\theta)$ 的无偏估计量，如果对所有 $\theta \in \Theta$，$\mathrm{E}_\theta T = g(\theta)$。偏差(bias)定义为 $\mathrm{E}_\theta T - g(\theta)$。

因此，$\mathrm{MSE}(\theta; T)$ 的第二项是偏差的平方。对于无偏估计量，这一项恒为 0。

这似乎是非常理想的，但并不总是如此。实际上，一定想满足无偏性条件, 这个统计量可能会方差非常大，从而在第一项中失去在第二项中获得的优势。通常，小方差会导致大偏差，而小偏差会导致大方差。因此，我们必须在两项之间权衡(tradeoff)。

估计量的标准差 $\sigma_\theta(T) = \sqrt{\mathrm{var}_\theta T}$ 也被称为标准误差。这不应与观察值的标准差混淆, 观察值的标准差是可以计算的, 但是估计量的标准差基于数据点(观察值) 这一由真参数控制的分布产生的随机变量。所以原则上，标准误差 $\sigma_\theta(T)$ 取决于未知参数 θ，因此它本身也是未知的。由于合理的估计量的偏差通常很小，标准误差通常可以提供关于估计量好不好的一个大致想法。估计量本身的同时往往也会提供标准误差的估计。这将在置信区间时候在此提到.

因此，我们寻找标准误差小且偏差小的估计量。

例 均匀分布

设 $X_1, \ldots, X_n$ 为独立同分布的 $\mathrm{U}[0, \theta]$ 随机变量。估计量 $2\bar{X}$ 是无偏的，因为对于所有 $\theta > 0$，

$$\mathrm{E}_\theta(2\bar{X}) = \frac{2}{n}\sum_{i=1}^{n} \mathrm{E}_\theta X_i = \frac{2}{n}\sum_{i=1}^{n} \frac{\theta}{2} = \theta$$

该估计量的均方误差为:

$$\mathrm{MSE}(\theta; 2\bar{X}) = 4\,\mathrm{var}_\theta\,\bar{X} = \frac{4}{n^2}\sum_{i=1}^{n} \mathrm{var}_\theta\, X_i = \frac{\theta^2}{3n}$$

估计量 $X_{(n)}$ 是有偏的，因为对于所有 $\theta > 0$，

$$\mathrm{E}_\theta X_{(n)} = \int_0^\theta x n x^{n-1} \frac{1}{\theta^n} dx = \frac{n}{n+1}\theta$$

这个积分很重要, 应该自己试一试.

尽管如此，（对于 n 不太小时）我们仍然更倾向于选择 $X_{(n)}$ 而非 $2\bar{X}$，因为该估计量具有较小的均方误差:

$$\begin{aligned}
\mathrm{MSE}\left(\theta; X_{(n)}\right) &= \mathrm{var}_\theta\, X_{(n)} + \left(\mathrm{E}_\theta X_{(n)} - \theta\right)^2 \\
&= \theta^2 \frac{n}{(n+2)(n+1)^2} + \theta^2\left(\frac{n}{n+1} - 1\right)^2 = \frac{2\theta^2}{(n+2)(n+1)}
\end{aligned}$$

我们可以通过乘以常数来抵消 $X_{(n)}$ 的偏差：估计量 $(n+1)/nX_{(n)}$ 是 θ 的无偏估计量。然而，有偏估计量 $(n+2)/(n+1)X_{(n)}$ 比我们迄今提到的所有估计量都要好，因为

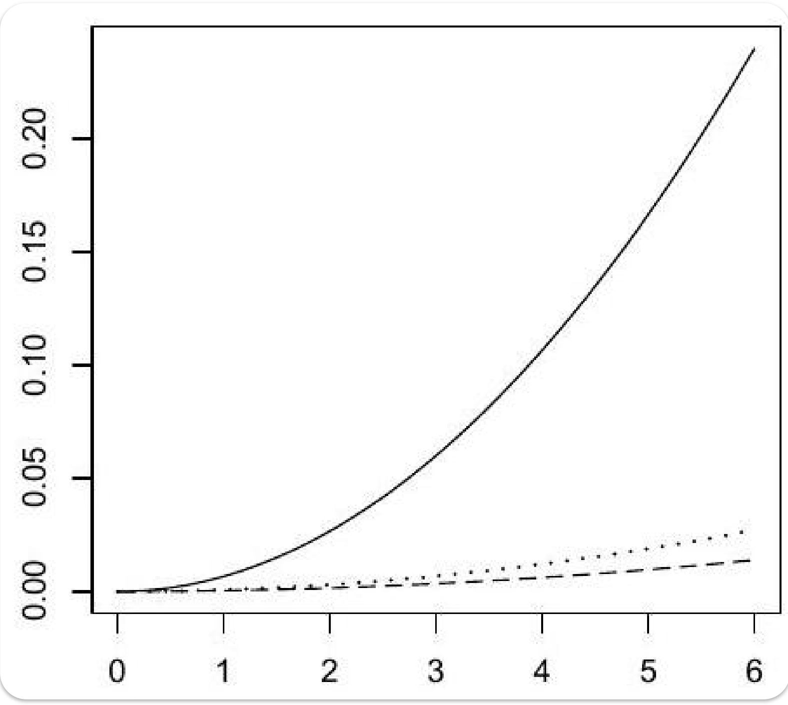
$$\mathrm{MSE}\left(\theta; \frac{n+2}{n+1}X_{(n)}\right) = \frac{\theta^2}{(n+1)^2}$$

下图显示了这个估计量的均方误差，以及 $X_{(n)}$ 和 $2\bar{X}$ 的均方误差，作为参数 θ 的函数，$n=50$。当 θ 值接近 0 时，$2\bar{X}$ 与其他两个估计量的均方误差差异较小，但当 θ 增大时，这些差异迅速增加。

仔细观察发现，对于不太小的 n 值，$(n+2)/(n+1)X_{(n)}$ 和 $X_{(n)}$ 的均方误差差异很小。然而，随着 n 的增大，$(n+2)/(n+1)X_{(n)}$ 相对于 $2\bar{X}$ 的精度提高迅速显现，因为前者的均方误差小了一个数量级。

我们已经指出，估计量 $(n+2)/(n+1)X_{(n)}$ 并不总是比 $2\bar{X}$ 在每个样本上给出更好的结果。

事实上，虽然我们知道 $\mathrm{MSE}(1;(n+2)/(n+1)X_{(n)}) < \mathrm{MSE}(1;2\bar{X})$
但这个严格不等式并不排除"并不是在每个样本上都给出更好的结果"，因为均方误差是一个期望值，可以视为大量实现的平均值。平均值可以为负而不必所有项都为负。不过我们可以从中得到一个结论<u>平均而言</u>，$(n+2)/(n+1)X_{(n)}$ 要（好得）多。



图估计量 $2\bar{X}$（实线）、$X_{(n)}$（虚线）和 $(n+2)/(n+1)X_{(n)}$（虚线）的均方误差作为 $\mathrm{U}[0,\theta]$ 参数的函数，$n=50$。

例 样本均值和样本方差

设 $X_1,\ldots,X_n$ 是独立同分布的随机变量，且其边际分布未知。我们希望估计这些观测值的期望 μ 和方差 σ^2。形式上，我们可以将 θ 设为一个未知的分布，即所谓的“非参数模型”，该模型不指定底层分布。参数 μ 和 σ^2 是底层分布的函数。

样本均值 $\bar{X}$ 和样本方差 S_X^2 定义为：

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad S_X^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$$

样本均值是 μ 的无偏估计量，因为

$$\mathrm{E}_\theta\bar{X}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}_\theta X_i=\mu$$

该估计量的均方误差为：

$$\mathrm{MSE}(\theta;\bar{X})=\mathrm{var}_\theta\,\bar{X}=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}_\theta\,X_i=\frac{\sigma^2}{n}$$

因此，样本均值的均方误差比基于单个观测值的估计量 X_i 的均方误差 $\mathrm{MSE}(\theta,X_i)=\mathrm{var}_\theta\,X_i=\sigma^2$ 小 n 倍。由于均方误差是估计的平方距离，我们可以得出结论，估计量 $\bar{X}$ 的质量提高了 $\sqrt{n}$ 倍。也就是说，要让估计量精度提高两倍，需要四倍的观测量。

样本方差是 σ^2 的无偏估计量，因为

$$\begin{aligned}\mathrm{E}_\theta S_X^2&=\mathrm{E}_\theta\frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i-\mu)+(\mu-\bar{X})\right]^2\\&=\mathrm{E}_\theta\frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i-\mu)^2+(\mu-\bar{X})^2+2(\mu-\bar{X})(X_i-\mu)\right]\\&=\frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}_\theta(X_i-\mu)^2-\frac{n}{n-1}\mathrm{E}_\theta(\bar{X}-\mu)^2=\sigma^2\end{aligned}$$

最后一个等号来自 $\mathrm{E}_\theta(X_i-\mu)^2=\mathrm{var}_\theta\,X_i=\sigma^2$ 和 $\mathrm{E}_\theta(\bar{X}-\mu)^2=\mathrm{var}_\theta\,\bar{X}=\sigma^2/n$。

通过进一步的计算，S_X^2 的均方误差可以表示为观测值的四阶矩，但我们在此不作讨论。

假设我们要寻找 μ^2 的无偏估计量。由于 $\bar{X}$ 是 μ 的无偏估计量，首先可以考虑使用 $\bar{X}^2$ 作为 μ^2 的估计量。**然而，这个估计量是有偏的：**

$$\mathrm{E}_\theta(\bar{X})^2=\mathrm{var}_\theta\,\bar{X}+\left(\mathrm{E}_\theta\bar{X}\right)^2=\frac{\sigma^2}{n}+\mu^2$$

因此很明显 $\mathrm{E}_\theta(\bar{X}^2-\sigma^2/n)=\mu^2$，但由于 σ^2 是未知参数，$\bar{X}^2-\sigma^2/n$ 不能作为估计量。**如果我们用无偏估计量 S_X^2 替换 σ^2，则可以得到 $\bar{X}^2-S_X^2/n$ 是 μ^2 的无偏估计量。(这个结论非常重要.)

例 样本理论

假设一个人口中有比例为 p 的个体具有某种特征 A。我们将比较三种估计 p 的方法：基于有放回抽样、无放回抽样和分层抽样。

在第一种方法中，我们从总体中抽取一个大小为 n 的样本，有放回地抽样，并使用样本中具有特征 A 的人数比例 X/n 来估计 p，其中 X 是样本中具有特征 A 的人数。此时，X 服从 $\mathrm{bin}(n,p)$ 分布，其期望值为 np，方差为 $np(1-p)$。由于 $\mathrm{E}_p(X/n)=p$ 对所有 p 成立，因此估计量 X/n 是无偏的。其均方误差为：

$$\mathrm{MSE}\left(p;\frac{X}{n}\right)=\mathrm{var}_p\left(\frac{X}{n}\right)=\frac{p(1-p)}{n}$$

根据二项分布我们能发现,这意味着，当 $p \approx 0$ 或 $p \approx 1$ 时，估计量表现更好，而当 $p = \frac{1}{2}$ 时表现最差。均方误差不依赖于总体的大小。通过选择足够大的样本(样本是从整体中抽取的)量，例如 $n \geq 1000$，我们可以获得均方误差至多为 $(1/4)/1000 = 1/4000$ 的估计量，无论总体大小是 800 还是几万亿人。

在第二种方法中，我们从总体中抽取一个大小为 n 的样本，无放回地抽样，并使用比例 Y/n 来估计 p，其中 Y 是样本中具有特征 A 的人数。此时，Y 服从 $\mathrm{hyp}(N, pN, n)$ 分布，其期望值为 np，方差为 $np(1-p)(N-n)/(N-1)$。因此，估计量 Y/n 也是无偏的，其均方误差为:

$$\mathrm{MSE}\left(p; \frac{Y}{n}\right) = \mathrm{var}_p\left(\frac{Y}{n}\right) = \frac{p(1-p)}{n}\frac{N-n}{N-1}$$

该均方误差小于 $\mathrm{MSE}(p; X/n)$，尽管对于 $n \ll N$ 时差异可以忽略不计。
这个"不放回比不放回的效果更好"并不奇怪：研究已经被研究过的人没有意义，如果不放回在统计上显然更有意义,但如果 $n \ll N$，这种情况发生的概率可以忽略不计。

在第三种方法中，我们首先将总体划分为若干个子群体，称为层。划分标准可以是地区、性别、年龄、收入、职业或其他背景变量。假设整个总体的规模为 N，而各个子群体的规模分别为 $N_1, \ldots, N_m$。为了方便，我们从第 j 个子群体中无放回地抽取 $(N_j/N)n$ 个人，形成一个分层样本，并使用 Z/n 估计 p，其中 Z 是样本中具有特征 A 的总人数。因此，$Z = Z_1 + \cdots + Z_m$，其中 Z_j 是从第 j 个子群体中抽取的具有特征 A 的人数。现在，$Z_1, \ldots, Z_m$ 是独立的，分别服从 $\mathrm{bin}\left((N_j/N)n, p_j\right)$ 分布，其中 p_j 是第 j 个子群体中具有特征 A 的比例。于是,

$$\begin{aligned}
\mathrm{E}_p\left(\frac{Z}{n}\right) &= \frac{1}{n}\sum_{j=1}^{m}\mathrm{E}_p Z_j = \frac{1}{n}\sum_{j=1}^{m}\frac{N_j}{N}np_j = \frac{1}{N}\sum_{j=1}^{m}N_j p_j = p \\
\mathrm{MSE}\left(p; \frac{Z}{n}\right) &= \mathrm{var}_p\left(\frac{Z}{n}\right) = \frac{1}{n^2}\sum_{j=1}^{m}\mathrm{var}_p Z_j \\
&= \frac{1}{n^2}\sum_{j=1}^{m}\frac{N_j n}{N}p_j(1-p_j) = \frac{p(1-p)}{n} - \frac{1}{n}\sum_{j=1}^{m}\frac{N_j}{N}(p_j - p)^2
\end{aligned}$$

因此，估计量 Z/n 也是无偏的，并且它的均方误差小于或等于 X/n 的均方误差。当 p_j 之间的差异较大时，这种差异尤为重要。分层抽样通常是首选的方法，尽管在实践中它可能意味着更多的工作。

类似的结果也适用于无放回抽样，只要层的大小和样本满足某些条件。然而，在这种情况下，分层并不总是带来更高的精度。