大数定律的应用——自助置信区间

假设 $X_1, X_2, \cdots, X_n$ 是一个来自分布函数为 $F(X;\theta)$ 的简单随机样本，θ 是我们感兴趣的总体参数，$T_n = g(X_1, X_2, \cdots, X_n)$ 为 θ 的估计量。我们知道，无论是对总体参数作估计还是检验，其核心关键的步骤就是确定 T_n 或关于 T_n 的函数的分布，或者至少要确定 T_n 或关于 T_n 的函数的期望方差等某些特征。只有分布或分布的某些特征确定，才能度量估计的性质好坏，评估检验的犯错概率。但这一核心步骤却是整个统计推断中最困难的一步。实际应用中，由于总体分布通常未知，T_n 的精确分布无法求得，即使我们已知总体分布的形式，有时 T_n 的精确分布也难于求得，而且就算我们能够推得 T_n 的精确分布，很多时候也因为形式复杂而难以付诸应用。

研究抽样分布（统计量或估计量的概率分布）的问题被 Fisher 认为是统计学的三大基本任务之一，然而直到今天，精确抽样分布依然大多来自于正态总体这个框架，且数量非常有限。为了解决抽样分布这一贯穿估计和检验的核心问题，人们依托大数定律、中心极限定理等许多极限理论又发展了许多渐近分布理论，从而求得了许多统计量的渐近分布。随着计算机的应用与普及，通过随机模拟方法获得近似抽样分布的方法迅速发展起来。$bootstrap$ 方法（自助法）便是其中之一，它是 $Efron$ 在 20 世纪 70 年代后期逐渐建立起来的以重抽样为核心的随机模拟方法，而它的理论支柱依然是以大数定律为核心的极限理论成果。自助法的基本步骤是：

1.自 $X_1, X_2, \cdots, X_n$ 的经验分布函数 $F_n(X;\theta)$ 中产生容量为 n 的简单随机样本（实际相当于对 $X_1, X_2, \cdots, X_n$ 作 n 次有放回随机抽样）$X_1^*, X_2^*, \cdots, X_n^*$，称其为自助样本；

2.依据上述自助样本计算 $T_n^* = g(X_1^*, X_2^*, \cdots, X_n^*)$ 的值；

3.重复步骤 1 和 2 B 次，得到 B 个自助统计量 $T_{n,1}^*, T_{n,2}^*, \cdots, T_{n,B}^*$.

4.求 B 个自助统计量的样本方差

$$v_{boot} = \frac{1}{B-1}\sum_{i=1}^{B}\left(T_{n,i}^* - \sum_{i=1}^{B}T_{n,i}^*\right)^2$$

则 v_{boot} 即为 T_n 方差的估计。

根据大数定律及相合估计的性质，我们有：当 $B \to \infty$ 时，

$$v_{boot} \xrightarrow{a.s.} Var_{F_n}[T_n]$$

因此我们可以用 v_{boot} 估计 $Var_{F_n}[T_n]$（注：具体来说，可以从两个角度获知上述收敛性是成立的：一方面，相合估计的函数是相应参数的同一函数的相合估计，由于样本方差可以表示为样本均值的函数，总体方差可以表示为总体期望的同一函数，并且根据大数定律有样本均值是总体期望的相合估计，因此样本方差也是总体方差的相合估计；另一方面，理论上可以证明：当样本为来自总体的简单随机样本时，样本方差的期望为总体方差，样本方差的方差 $\big(\mu_4 - (n-3)\sigma^4/(n-1)\big)/n$ 随着样本量趋于无穷而趋于 0，这里 μ_4 为总体的四阶中心矩，σ^4 为总体方差的平方。）又根据格里文科定理，即

$$\sup_x |F_n(X) - F(X)| \xrightarrow{a.s.} 0$$

因此我们可以用 $Var_{F_n}[T_n]$ 估计 $Var_F[T_n]$，综上，我们可以用 v_{boot} 估计 $Var_F[T_n]$.

自助置信区间的构造有如下几种方式：

1.自助正态置信区间（bootstrap normal confidence interval）为：$T_n \pm z_{\alpha/2}\sqrt{v_{boot}}$，其

中 $z_{\alpha/2}$ 是标准正态分布上 $(\alpha/2)$ 分位数，显然这种人为假定统计量服从正态分布而构造置信区间的方法在很多情况下是不可取的。

2.自助枢轴置信区间（bootstrap pivotal confidence interval）为：$\left(T_n-\left[q^*_{(1-\alpha/2)}-T_n\right],\right.$ $\left.T_n+\left[T_n-q^*_{(\alpha/2)}\right]\right)$，这里 q^*_β 为 $T^*_{n,1},T^*_{n,2},\cdots,T^*_{n,B}$ 的 β 分位数（注：其基本原理为，假设 $T_n-\theta$ 为一个枢轴量，故其分布函数 $D(x)$ 与参数 θ 无关，由此可得到 θ 的 $1-\alpha$ 置信区间为 $\left(T_n-D_{(1-\alpha/2)},T_n-D_{(\alpha/2)}\right)$，$D_{(m)}$ 为分布函数 $D(x)$ 的 m 分位数，它实际上等于 T_n 的 m 分位数减去 θ，自助枢轴置信区间正是这一枢轴置信区间的估计，它用 T_n 的自助分位数作为 T_n 分位数的估计，并用 T_n 作为 θ 的估计，即用 $q^*_{(1-\alpha/2)}-T_n$ 估计 $D_{(1-\alpha/2)}$，用 $q^*_{(\alpha/2)}-T_n$ 估计 $D_{(\alpha/2)}$）。

3.自助学生化枢轴区间（bootstrap studentized pivotal interval）为：$\left(T_n-z^*_{(1-\alpha/2)}\sqrt{v_{boot}},\right.$

$\left.T_n-z^*_{(\alpha/2)}\sqrt{v_{boot}}\right)$，这里 z^*_β 为 $Z^*_{n,1},Z^*_{n,2},\cdots,Z^*_{n,B}$ 的 β 分位数，$Z^*_{n,i}=\left(T^*_{n,i}-T_n\right)/\widehat{se}^*_i$，$\widehat{se}^*_i$ 为 $T^*_{n,i}$ 的标准误差的一个估计，通过对第 i 个自助样本 $X^*_{1,i},X^*_{2,i},\cdots,X^*_{n,i}$ 再次施行自助法可以得到 $\widehat{se}^*_i$，也可以应用自助法的非参数 delta 方法来计算 $\widehat{se}^*_i$ 并且按此方法计算所耗计算机资源要小得多。

4.自助分位数区间（bootstrap percentile interval）为：$\left(q^*_{(\alpha/2)},q^*_{(1-\alpha/2)}\right)$，这里 q^*_β 为 $T^*_{n,1},T^*_{n,2},\cdots,T^*_{n,B}$ 的 β 分位数。

理论上可以证明，上述四种自助置信区间中学生化枢轴区间最精确（注：区间长度最短）。上述介绍的自助方法属于非参数的范畴，如果我们预先知道了 $F(X;\theta)$ 的形式，只是不知道 θ 的取值，那么可以利用极大似然的方法首先估计出 θ，然后自 $F\left(X;\hat{\theta}\right)$ 产生自助样本，其后续步骤同非参数 $bootstrap$ 方法是一致的。

自助置信区间与传统置信区间在构造思路上是基本一致的，即找到参数的合适估计量，然后找到该估计量或估计量函数的抽样分布，最后依据该抽样分布构建参数的置信区间。它们的区别主要体现在寻找估计量抽样分布这一关键过程上，传统置信区间在寻找估计量的抽样分布时，通常是在已知总体分布、或对总体分布作出某些假定的前提下利用数学方法导出估计量的精确或渐近抽样分布，而自助置信区间在寻找估计量的抽样分布时，则是通过重抽样方法获得估计量的近似抽样分布。

需要指出的是，自助法并非一种放之四海而皆准的万能方法，它的合理性需要某些正则条件的保证，应此绝不能不加考虑地盲目应用。